CHANGE-POINT METHODS FOR OVERDISPERSED COUNT DATA

THESIS

Brian A. Wilken, Captain, United States Air Force

AFIT/GOR/ENS/07-26

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY
Wright-Patterson Air Force Base, Ohio

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the United States Government.

AFIT/GOR/ENS/07-26

CHANGE-POINT METHODS FOR OVERDISPERSED COUNT DATA

THESIS

Presented to the Faculty

Department of Operational Sciences

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Operations Research

Brian A. Wilken, B.A.

Captain, United States Air Force

March 2007

AFIT/GOR/ENS/07-26

CHANGE-POINT METHODS FOR OVERDISPERSED COUNT DATA

Brian A. Wilken, B.A.
Captain, United States Air Force

Approved:

_____	_____
Dr. Marcus B. Perry (Chairman) Assistant Professor of Operations Research	Date
_____	_____
Dr. Sharif Melouk (Member) Assistant Professor of Operations Research	Date

Abstract

A control chart is often used to detect a change in a process. Following a control chart signal, knowledge of the time and magnitude of the change would simplify the search for and identification of the assignable cause. In this research, emphasis is placed on count processes where overdispersion has occurred. Overdispersion is common in practice and occurs when the observed variance is larger than the theoretical variance of the assumed model. Although the Poisson model is often used to model count data, the two-parameter gamma-Poisson mixture parameterization of the negative binomial distribution is often a more adequate model for overdispersed count data. In this research effort, maximum likelihood estimators for the time of a step change in each of the parameters of the gamma-Poisson mixture model are derived. Monte Carlo simulation is used to evaluate the root mean square error performance of these estimators to determine their utility in estimating the change point, following a control chart signal. Results show that the estimators provide process engineers with accurate and useful estimates for the time of step change. In addition, an approach for estimating a confidence set for the process change point will be presented.

Dedication

To my wife and children

Acknowledgments

I would like to thank several people for their assistance throughout this research effort. First, I would like to thank my advisor, Dr. Marcus Perry. His guidance and feedback was crucial to my understanding and application of the ideas presented in this thesis. I am also grateful to my reader, Dr. Sharif Melouk. An additional thanks goes to my friends from class GOR-07M. I thank each of them for all their help and support throughout the program. Last, I must thank my wife and children for their loving support and sacrifices over the last 18 months. This program and research would not have been possible without them.

Brian A. Wilken

Table of Contents

Page

Abstract iv

Dedication v

Acknowledgments vi

List of Figures ix

List of Tables xiii

1. Introduction 1

   1.1 Review of Statistical Process Control 1
   1.2 Problem Definition 9
   1.3 Research Objectives and Assumptions 13
   1.4 Thesis Organization 14

2. Literature Review 16

   2.1 Overview of a Change-Point Model 16
   2.2 Change-Point Estimation 17
   2.3 Conclusion 22

3. Methodology 23

   3.1 Derivation of the gamma-Poisson Mixture Model 23
   3.2 Change-Point Model for a Step Change in Mean 27
   3.3 Derivation of the MLE When ν_a is Unknown 28
   3.4 Change-Point Model for a Step Change in the Overdispersion Parameter 32
   3.5 Derivation of the MLE When α_a is Unknown 33
   3.6 Newton's Method for Finding the MLE of α_a 35
   3.7 Confidence Sets 38
   3.8 Conclusion 40

4. Results and Analysis 41

   4.1 Monte Carlo Simulation When ν_a is Unknown 41
   4.2 Performance of the MLE for Step Change in Mean 43

4.3 Cardinality and Coverage Performances of Confidence Set Estimators for Step Change in Mean 50
4.4 Monte Carlo Simulation When α_a is Unknown 54
4.5 CUSUM Control Chart for Detecting a Change in Overdispersion Parameter 55
4.6 Performance of the MLE for Step Change in the Overdispersion Parameter ... 57
4.7 Cardinality and Coverage Performances of Confidence Set Estimators for Step Change in Overdispersion Parameter 60
4.8 Change-Point Analysis Applied to Iraq War Coalition Casualty Data 64
4.9 Conclusion 68

5. Conclusions and Future Research 70

5.1 Summary and Conclusions 70
5.2 Future Research 72

Appendix A: Simulation Data for Step Change in Mean 74

Appendix B: Cardinality and Coverage Surface Plots for Step Change in Mean 80

Appendix C: Simulation Data for Step Change in Overdispersion Parameter 86

Appendix D: Cardinality and Coverage Surface Plots for Step Change in Overdispersion Parameter 88

Bibliography 91

Vita 93

List of Figures

Page

Figure 1: Basic Control Chart 3

Figure 2: Overdispersed Count Data 11

Figure 3: Equidispersed Count Data 12

Figure 4: Newton's Method Coded in MATLAB 36

Figure 5: Plots Showing the Effect of Increasing α Over the Log-likelihood Function, and P and Q from MATLAB code in Figure 4 for Single Observation y. ν=5 and y=5. 37

Figure 6: Log-Likelihood Plot 39

Figure 7: Effect on RMS due to an Increasing α. 45

Figure 8: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Set Cardinality, Coverage, Percent Increase from ν_0, and Reference Value D. $\alpha = 1$, ν_0=1, τ=50, and $N = 10,000$. 52

Figure 9: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Set Cardinality, Coverage, Percent Increase from ν_0, and Reference Value D. $\alpha = 5$, ν_0=1, τ=50, and $N = 10,000$. 52

Figure 10: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Cardinality, Coverage, Percent Increase from ν_0, and Reference Value D. $\alpha = 10$, ν_0=1, τ=50, and $N = 10,000$. 53

Figure 11: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Cardinality, Coverage, Percent Increase from ν_0, and Reference Value D. $\alpha = 50$, ν_0=1, τ=50, and $N = 10,000$. 53

Figure 12: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Cardinality, Coverage, Percent Decrease from α_0, and Reference Value D. ν=5, α_0=5, τ=50, and $N = 10,000$. 62

Figure 13: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Cardinality, Coverage, Percent Decrease from α_0, and Reference Value D. ν=20, α_0=5, τ=50, and $N = 10,000$. 62

Figure 14: Control Chart, Iraq Coalition Casualties 07/19/03 – 10/26/03..... 66

Figure 15: Log-Likelihood Plot of Iraq War Casualty Data 67

Figure 16: Coalition Casualties 07/19/03 – 12/31/03 67

Figure 17: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Set Cardinality, Coverage, Percent Increase from ν_0, and Reference Value D. $\alpha = 1$, ν_0=1, τ=50, and $N = 10,000$. 80

Figure 18: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Set Cardinality, Coverage, Percent Increase from ν_0, and Reference Value D. $\alpha = 1$, ν_0=5, τ=50, and $N = 10,000$. 80

Figure 19: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Set Cardinality, Coverage, Percent Increase from ν_0, and Reference Value D. $\alpha = 1$, ν_0=20, τ=50, and $N = 10,000$. 81

Figure 20: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Set Cardinality, Coverage, Percent Increase from ν_0, and Reference Value D. $\alpha = 5$, ν_0=1, τ=50, and $N = 10,000$. 81

Figure 21: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Set Cardinality, Coverage, Percent Increase from ν_0, and Reference Value D. $\alpha = 5$, ν_0=5, τ=50, and $N = 10,000$. 82

Figure 22: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Set Cardinality, Coverage, Percent Increase from ν_0, and Reference Value D. $\alpha = 5$, ν_0=20, τ=50, and $N = 10,000$. 82

Figure 23: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Set Cardinality, Coverage, Percent Increase from ν_0, and Reference Value D. $\alpha = 10$, ν_0=1, τ=50, and $N = 10,000$. 83

Figure 24: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Set Cardinality, Coverage, Percent Increase from ν_0, and Reference Value D. $\alpha = 10$, ν_0=5, τ=50, and $N = 10,000$. 83

Figure 25: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Set Cardinality, Coverage, Percent Increase from ν_0, and Reference Value D. $\alpha = 10$, ν_0=20, τ=50, and $N = 10,000$. 84

Figure 26: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Set Cardinality, Coverage, Percent Increase from ν_0, and Reference Value D. $\alpha = 50$, ν_0=1, τ=50, and $N = 10,000$. 84

Figure 27: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Set Cardinality, Coverage, Percent Increase from ν_0, and Reference Value D. $\alpha = 50$, ν_0=5, τ=50, and $N = 10,000$. 85

Figure 28: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Set Cardinality, Coverage, Percent Increase from ν_0, and Reference Value D. $\alpha = 50$, ν_0=20, τ=50, and $N = 10,000$. 85

Figure 29: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Cardinality, Coverage, Percent Decrease from α_0, and Reference Value D. ν=5, α_0=1, τ=50, and $N = 10,000$. 88

Figure 30: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Cardinality, Coverage, Percent Decrease from α_0, and Reference Value D. ν=5, α_0=5, τ=50, and $N = 10,000$. 88

Figure 31: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Cardinality, Coverage, Percent Decrease from α_0, and Reference Value D. ν=5, α_0=10, τ=50, and $N = 10,000$. 89

Figure 32: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Cardinality, Coverage, Percent Decrease from α_0, and Reference Value D. ν=20, α_0=1, τ=50, and $N = 10,000$. 89

Figure 33: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Cardinality, Coverage, Percent Decrease from α_0, and Reference Value D. ν=20, α_0=5, τ=50, and $N = 10,000$. 90

Figure 34: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Cardinality, Coverage, Percent Decrease from α_0, and Reference Value D. ν=20, α_0=10, τ=50, and $N = 10,000$. 90

List of Tables

Page

Table 1: Observed Count at Each Observation i 30

Table 2: Log-Likelihood Values at each Potential Change Point 32

Table 3: Performance of MLE for α=1, $\nu_0 = 5$, τ=50, and $N = 10,000$ 44

Table 4: Effect on MLE for Increasing α 45

Table 5: Comparison of MLE for $\alpha = 1$ and α=50 46

Table 6: Performance of MLE for α=5, $\nu_0 = 5$, τ=50, and $N = 10,000$ 46

Table 7: Performance of MLE for α=10, $\nu_0 = 5$, τ=50, and $N = 10,000$ 47

Table 8: Performance of MLE for α=50, $\nu_0 = 5$, τ=50, and $N = 10,000$ 47

Table 9: Comparison of MLE for $\alpha = 50$ for $\nu_0 = 1,\ 5$, and 10. 48

Table 10: Performance of MLE for α=50, $\nu_0 = 1$, τ=50, and $N = 10,000$ 49

Table 11: Performance of MLE for α=50, $\nu_0 = 20$, τ=50, and $N = 10,000$ 49

Table 12: Effect of the Change in τ on Coverage Probability.
$\alpha = 5,\ \nu_0 = 20,$ and $N = 10,000$ 54

Table 13: Effect of the Change in τ on Average Size of Confidence Sets.
$\alpha = 5,\ \nu_0 = 20,$ and $N = 10,000$ 54

Table 14: Performance of MLE for $\nu = 5,\ \alpha_0 = 10,\ \tau = 50,$ and N=10,000 57

Table 15: Comparison of MLE for $\nu = 5$ and ν=20 for Various Decreases in α Relative to α_0. $\alpha_0 = 10$, τ=50, and $N = 10,000$ 58

Table 16: Comparison of MLE for $\nu = 20$ for $\alpha_0 = 1,\ 5,$ and 10. 59

Table 17: Performance of MLE for $\nu = 20,\ \alpha_0 = 1,\ \tau = 50,$ and N=10,000 59

Table 18: Performance of MLE for $\nu = 20,\ \alpha_0 = 5,\ \tau = 50,$ and N=10,000 59

Table 19: Performance of MLE for $\nu = 20,\ \alpha_0 = 10,\ \tau = 50,$ and N=10,000 60

Table 20: Effect of the Change in τ on Coverage Probability.
$\alpha = 5,\ \nu_0 = 20,$ and $N = 10,000$ 63

Table 21: Effect of the Change in τ on Average Size of Confidence Sets.
$\alpha = 5,\ \nu_0 = 20,$ and $N = 10,000$ 63

Table 22: Performance of MLE for $\alpha = 1,\ \nu_0 = 1,\ \tau = 50,$ and $N = 10,000$ 74

Table 23: Performance of MLE for $\alpha = 1,\ \nu_0 = 5,\ \tau = 50,$ and $N = 10,000$ 74

Table 24: Performance of MLE for $\alpha = 1,\ \nu_0 = 20,\ \tau = 50,$ and $N = 10,000$ 75

Table 25: Performance of MLE for $\alpha = 5,\ \nu_0 = 1,\ \tau = 50,$ and $N = 10,000$ 75

Table 26: Performance of MLE for $\alpha = 5,\ \nu_0 = 5,\ \tau = 50,$ and $N = 10,000$ 76

Table 27: Performance of MLE for $\alpha = 5,\ \nu_0 = 20,\ \tau = 50,$ and $N = 10,000$ 76

Table 28: Performance of MLE for $\alpha = 10,\ \nu_0 = 1,\ \tau = 50,$ and $N = 10,000$ 77

Table 29: Performance of MLE for $\alpha = 10,\ \nu_0 = 5,\ \tau = 50,$ and $N = 10,000$ 77

Table 30: Performance of MLE for $\alpha = 10,\ \nu_0 = 20,\ \tau = 50,$ and $N = 10,000$ 78

Table 31: Performance of MLE for $\alpha = 50,\ \nu_0 = 1,\ \tau = 50,$ and $N = 10,000$ 78

Table 32: Performance of MLE for $\alpha = 50,\ \nu_0 = 5,\ \tau = 50,$ and $N = 10,000$ 79

Table 33: Performance of MLE for $\alpha = 50,\ \nu_0 = 20,\ \tau = 50,$ and $N = 10,000$ 79

Table 34: Performance of MLE for $\nu = 5,\ \alpha_0 = 1,\ \tau$=50, and $N = 10,000$ 86

Table 35: Performance of MLE for $\nu = 5,\ \alpha_0 = 5,\ \tau$=50, and $N = 10,000$ 86

Table 36: Performance of MLE for $\nu = 5,\ \alpha_0 = 10,\ \tau$=50, and $N = 10,000$ 86

Table 37: Performance of MLE for $\nu = 20$, $\alpha_0 = 1$, τ=50, and $N = 10,000$ 87

Table 38: Performance of MLE for $\nu = 20$, $\alpha_0 = 5$, τ=50, and $N = 10,000$ 87

Table 39: Performance of MLE for $\nu = 20$, $\alpha_0 = 10$, τ=50, and $N = 10,000$ 87

CHANGE-POINT METHODS FOR OVERDISPERSED COUNT DATA

1. Introduction

This chapter will start with some relevant background information on statistical process control (SPC) that sets the foundation for this research effort. Following the background review, the problem definition will be introduced along with the research objectives and assumptions of this thesis. The last section of this chapter will lay out the organization of this thesis.

1.1 Review of Statistical Process Control

The focus of this research is on change-point estimation within an SPC context. Before going further with this research, a brief overview of SPC topics and terminology pertaining to this research is reviewed. Refer to Montgomery [9] and Ryan [19] for more detailed descriptions of SPC and its applications.

1.1.1 SPC Terminology and Concepts

Montgomery [9] defines statistical process control as a set of problem-solving tools that can be applied to a process to achieve stability and improve capability by reducing variability. This term, *process*, is usually thought of in a manufacturing setting, but can include non-manufacturing settings as well. Non-manufacturing processes can include tracking errors in some paperwork process of a business, or even monitoring

daily coalition casualties in the Iraq war, which is examined in Chapter 4. Of the different SPC tools available, such as histograms and cause-and-effect diagrams, this research will only focus on the control chart, specifically the Shewhart control chart and the cumulative sum control chart. Before getting into the specifics of each of these control charts, some overall SPC terminology needs to be addressed.

Any process, no matter how finely tuned, will always contain some variability. Natural variability, or uncontrollable variability, is inherent in any process. A process operating in a state only affected by natural variability is said to be statistically *in-control*. However, other kinds of variability may be present in the output of a process. For instance, in a manufacturing context these non-natural sources of variability, or controllable sources of variability, may be introduced into a process by improperly calibrated machines, defective raw materials, or even human error. Causes of these sources of variability are referred to as *assignable causes*. A process is said to be *out-of-control* if assignable causes are present. Thus, one of the primary objectives of SPC is to quickly detect a change in the process, look for assignable causes, and take corrective action to bring the process back to a state of statistical control as quickly as possible.

One way a process can be monitored for changes is through the use of a control chart. A basic Shewhart control chart can be seen in Figure 1. The *center line* represents the expected in-control value of the quality characteristic being measured. Examples of quality characteristics include numerical measurements like temperature or diameter, or frequency counts, such as the number of defects inspected per item. The upper control limit (UCL) and lower control limit (LCL) are also shown. The control limits are chosen in such a way that if the process is statistically in-control, then the rate at which the chart

falsely signals is controlled at some desired level. In Figure 1, the observations are connected by a line to show how the observations evolve over time.

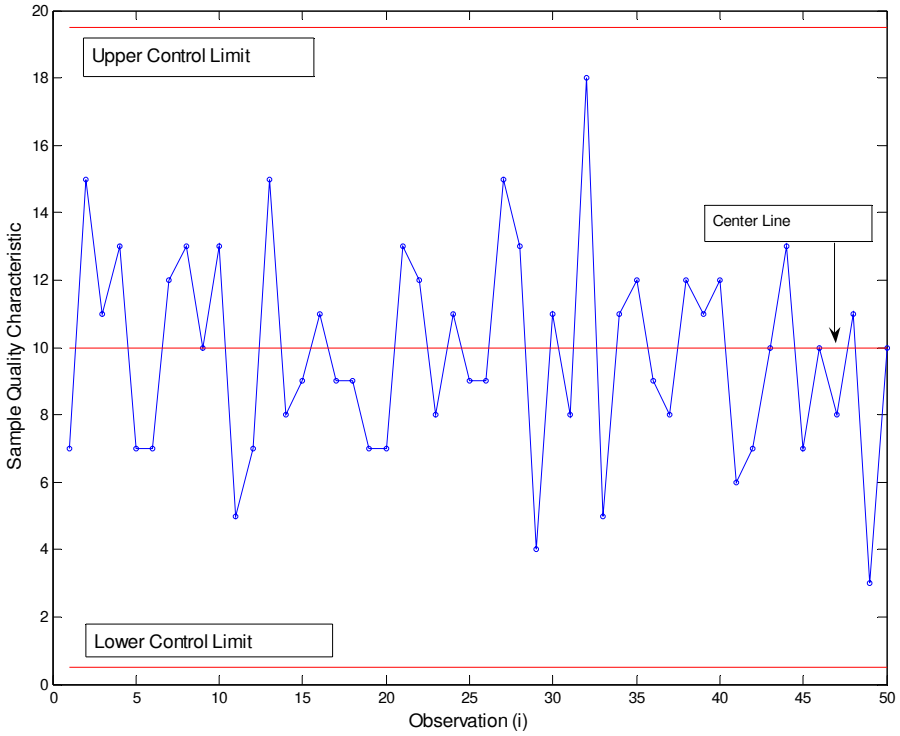

Figure 1: Basic Control Chart

Control charts are closely related to hypothesis testing. The control chart continually tests the null hypothesis that the process is statistically in-control. If the observations stay within the control limits, this is equivalent to failing to reject the null hypothesis. Likewise, a point plotted outside the control limits is equivalent to rejecting the null hypothesis. For example, Figure 1 may represent some process where the hypothesized mean of the quality characteristic being measured is 10 units. The hypothesis test is

$$H_0 : \mu_i = 10; \quad i = 1, 2, ..., T$$
$$H_1 : \mu_i \neq 10; \quad i = 1, 2, ..., T.$$

Looking at Figure 1, one can fail to reject the null hypothesis, which means that there is not enough evidence to conclude that the process is out-of-control. Wackerly, Mendenhall, and Scheaffer [22] provide an excellent overview of mathematical statistics, including hypothesis testing.

1.1.2 Shewhart Control Chart

The first control chart used in this thesis is the Shewhart control chart, named after Walter Shewhart, who developed these ideas while working at Bell Labs during the 1920's. A general model used to construct a Shewhart control chart is as follows. Let *m* denote a sample statistic measuring some quality characteristic of interest. The mean of *m* is μ_m and the standard deviation of *m* is σ_m. The center line, upper control limit, and lower control limit are

$$\begin{aligned} UCL &= \mu_m + L\sigma_m \\ CenterLine &= \mu_m \\ LCL &= \mu_m - L\sigma_m \end{aligned} \tag{1.1}$$

where *L* denotes the distance of the control limits from the center line expressed in standard deviation units. A control chart designed according to these principles is a Shewhart control chart, and using *L* = 3 creates a *three-sigma* (3σ) control chart.

Integrating all of the above concepts and terminology, an SPC process monitored by a 3σ control chart can be described as follows. In *phase 1*, process data are gathered over some period of time in which the process is assumed to be in-control. The data are analyzed, trial control limits are established using equation (1.1), and process engineers

then use the control chart to bring the process into a state of statistical control. Once confident the process is in control and the actual 3σ control limits of the in-control process have been calibrated, then *phase II* is implemented. During phase II, the process is monitored by comparing the control chart statistic computed from each sample drawn to the control limits. If the process shifts to an out-of-control state, the control chart will likely signal, and the process engineers can begin their search for any assignable causes that may have caused the process to change. Their ultimate goal is to make appropriate process adjustments to get the process back into a state of statistical control as quickly as possible.

One way to evaluate the performance of a control chart implemented in phase II monitoring is to look at the Average Run Length (ARL). ARL is defined as the number of observations plotted on the control chart before a point plots outside the control limits, signaling an out-of-control condition. For a Shewhart control chart where the observations are uncorrelated, the ARL is calculated from

$$\text{ARL} = \frac{1}{p} \tag{1.2}$$

where p is the probability any point exceeds the control limits. For a 3σ control chart where the mean is the quality characteristic of interest, $p = 0.0027$ is the probability that a point falls outside the control limits. The average run length of the chart when the process is in-control (ARL_0) is

$$\text{ARL}_0 = \frac{1}{0.0027} = 370.$$

The interpretation is even if the process remains in-control, an out-of-control signal will be generated on average, every 370 samples. Obviously a large ARL_0 is a desirable property of a control chart.

Shewhart control charts are designed a little differently depending on the quality characteristic being measured. The two main types of Shewhart control charts are control charts for *variables* and *attributes*. If the quality characteristic being measured can be represented on a numerical scale, such as temperature or length, the numerical scale is called *variable*. The two main control charts for variables are $\bar{x}$ and R charts (or S charts) used to monitor the mean and variance of numerical quality characteristics, respectively. However, not all quality characteristics are represented numerically. Perhaps each item inspected is classified as conforming (nondefective) or nonconforming (defective). Quality characteristics of this type are called *attributes*. There are three main Shewhart-type control charts for attributes:

1. *p*-charts for binomial counts. Monitors the fraction of nonconforming product produced by a process.
2. *c*-charts for Poisson counts. Monitors the number of nonconformities observed.
3. *u*-charts for Poisson counts. Monitors the average number of nonconformities observed per unit.

Again, the reader is referred to Montgomery [9] and Ryan [19] for more detailed descriptions of SPC, including the details of these control charts.

1.1.3 Cumulative Sum Control Chart

The Shewhart control chart is easy to implement, widely used, and good for detecting large changes in a process. It is particularly useful during the phase I implementation of a process, where the process is not finely tuned yet. Shewhart control charts have the disadvantage of only using information about the process contained in the last observation while ignoring the previous sequence of observations. Cumulative sum (CUSUM) control charts are effective alternatives for detecting small changes, using information from the current and previous sequence of observations. Therefore, CUSUMs are good choices to implement during the phase II monitoring of a process when the in-control process parameters are assumed to be known or sufficiently estimated. Refer to Hawkins and Olwell [6] for a complete reference on CUSUM charts, including more detail regarding the theoretical foundations of the CUSUM discussed below.

The *tabular* CUSUM plots the statistics

$$C_i^{\pm} = \max(0,\ \pm(y_i - k) + C_{i-1}^{\pm}) \tag{1.3}$$

where $C_i^{\pm}$ is the cumulative sum at time *i*. The algorithm cumulates the difference between an observed value y_i and a reference value *k*. If the test statistic C_i^{+} (C_i^{-}) exceeds a *decision interval* h^{+} (h^{-}), the chart signals that an increase (decrease) has been detected, suggesting the process is out-of-control. The values of $h^{\pm}$ and $k^{\pm}$ are selected based on desired ARL properties.

The CUSUM statistic above is derived from the sequential probability ratio test (SPRT), which uses observed data sequentially, employing each observation y_i as it

becomes available. For each new *T*, the SPRT sequentially tests the null hypothesis $H_0 : \Theta_i = \Theta_0$ i = 1, 2, ..., T against the alterative hypothesis $H_a : \Theta_i = \Theta_a$ i = 1, 2, ..., T. Θ_0 and Θ_a are the in-control and pre-specified out-of-control values of the process parameters, respectively. Each hypothesis is associated with a PDF $f_0(y_i \mid \Theta_0)$ and $f_a(y_i \mid \Theta_a)$ respectively.

In SPRT, the likelihood ratio is given by $Z_i^{\pm} = \prod_{i=1}^{T} \frac{f_a(y_i|\Theta_a)}{f_0(y_i|\Theta_0)}$, while the log-likelihood ratio is defined as $C_i^{\pm} = \sum_{i=1}^{T} \ln(\frac{f_a(y_i|\Theta_a)}{f_0(y_i|\Theta_0)})$. The SPRT operates by comparing $C_i^{\pm}$ to a decision interval $h^{\pm}$ at each new observation. New samples are collected until $C_i^{\pm} > h^{\pm}$; the point that the test concludes in favor of H_a. In the CUSUM approach, the null hypothesis of in-control is never accepted. Hence, the test is restarted each time the evidence favors the null hypothesis. Thus, samples are obtained until the null hypothesis is rejected in favor of the alternative hypothesis.

It is worth noting the tabular CUSUM algorithm uses counters (or built-in change-point estimators) that suggest when a process shift occurred. For example, the quantity N^+ indicates the number of consecutive periods or observations that the CUSUM C_i^+ has been nonzero. As an example, suppose that the control chart signaled at period T = 20, with N^+ = 5. This would indicate that five consecutive periods have elapsed since C_i^+ rose above the value of zero. Thus, it is likely the process was last in-control at period 20 - 5 = 15, and that the shift likely occurred between periods 15 and 16. After reviewing the basics of SPC and learning how a 3σ control chart and CUSUM chart are used to detect changes in a process, the problem definition can now be defined.

1.2 Problem Definition

It was discussed above that a control chart is often used to detect a change in a process. Following a control chart signal, knowledge of the time and magnitude of the change would simplify the search for and identification of the assignable cause. In this research, emphasis is placed on count processes where the data are overdispersed, which implies that the observed variance is greater than the hypothesized variance. This section will first discuss overdispersion in general. Then, some discussion is provided that addresses the problem of estimating a process change point where the observations are overdispersed count data.

1.2.1 Overdispersion

Consider a modeling scenario where the quality characteristic of interest is a count. An event count is the realization of a non-negative integer-valued random variable [4]. Examples include the number of war casualties in Iraq per day, the number of defects a product manufacturer experiences per week, or the number of accidents a construction company observes per day. It is well known that the Poisson probability distribution provides a reasonable probability distribution to model count data. See Haight [5] for a comprehensive reference on the Poisson distribution.

Let Y_i denote a random variable distributed Poisson(λ_i), where $\lambda_i > 0$ is the rate parameter at time *i* and denotes the expected value of Y_i. The Poisson probability mass function (PMF), a discrete function defined for $y_i \in [0,\ 1,\ 2,\ ...]$, is

$$P(Y_i = y_i) = f(y_i \mid \lambda_i) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} \tag{1.4}$$

where, $P(Y_i = y_i \mid \lambda_i)$ denotes the probability that the random variable Y_i will take on the value y_i, given λ_i.

Suppose it is postulated that $\lambda_i = \lambda_0$ for all $i = 1, 2, ..., T$. Under Poisson model assumptions, one should expect that $E(Y_i) = Var(Y_i) = \lambda_0$, i.e., *equidispersion*, for all $i = 1, 2, ..., T$. Unfortunately, this is rarely the case in practice. More often the observed variance of a random sample is greater than the theoretical variance, implying that the count data are overdispersed. If Poisson count data are collected over time in discrete samples, overdispersion can occur as a result of uncontrollable fluctuations in the rate parameter over this time domain.

For example, consider Figure 2. The solid line represents a vector of λ_i's, where the i^{th} element is the value of the rate parameter for the i^{th} sample. This vector is generally not observable; instead, only the noisy counts can be observed (represented by circles in Figure 2).

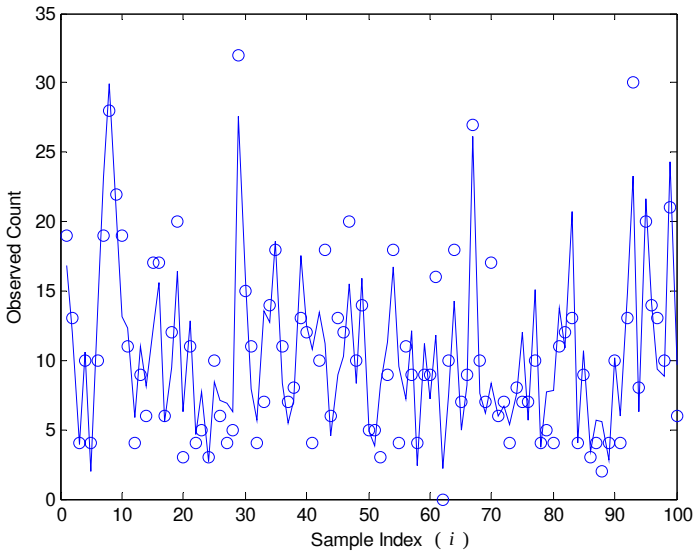

Figure 2: Overdispersed Count Data

Under the postulated model that $\lambda_i = \lambda_0$ for all $i = 1, 2, ..., T$, the sample mean and sample variance of the 100 observations should be nearly equal. The estimate of the mean of the data in Figure 2 is given by $\hat{\lambda}_0 = 10.40$, whereas the variance is estimated to be $\hat{\sigma}_y^2 = 41.58$. This overdispersion is a direct result of the fluctuations in the λ_i's over the sampling interval. If the rate parameter remains constant over the sampling interval, then the data are no longer overdispersed, as shown in Figure 3. The estimate for the mean of the count data in Figure 3 is given by $\hat{\lambda}_0 = 9.48$, whereas the variance is estimated to be $\hat{\sigma}_y^2 = 9.57$. Clearly, the mean of the data is much closer to the variance of the data given in Figure 3.

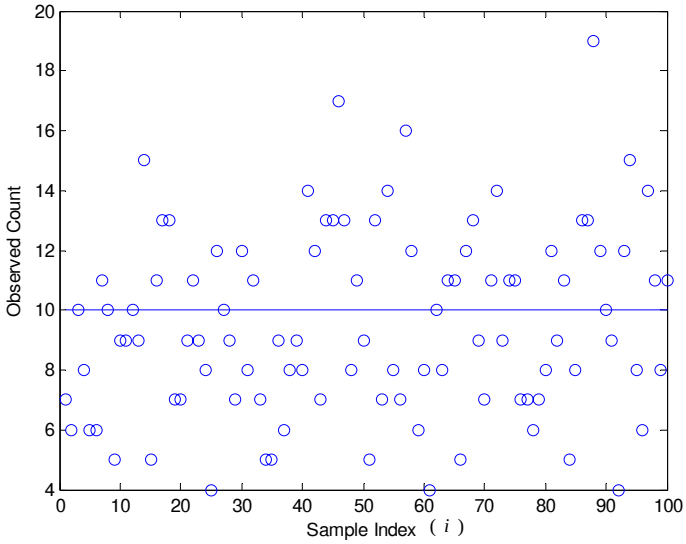

Figure 3: Equidispersed Count Data

It was discussed above that the Poisson probability distribution is often not an adequate means to model overdispersed count data. Again, this is due to the observed variance of a random sample being greater than the theoretical variance. Thus, one objective of this research is to derive a probability distribution that will adequately account for overdispersion in count data that is otherwise distributed as Poisson. In this section overdispersion was defined. It is shown in Chapter 3 that the two-parameter gamma-Poisson mixture model is a reasonable model for overdispersed count data. The next section will discuss the problem of change-point estimation within an SPC setting, where the count data are assumed to be overdispersed.

1.2.2 Change Point Estimation

Often the process engineer would like to not only detect a change in the process, but also estimate when the change actually occurred. Knowledge of the time and magnitude of the change would simplify the search for and identification of the

assignable cause. Having an estimate for the change point can reduce the costs associated with misdiagnosing a control chart signal. These costs include the time and resources used to track down any assignable causes, as well as any costs associated with unnecessary process adjustments. Thus, the primary focus of this research is to derive and evaluate maximum likelihood change-point estimators (MLEs) for the time of a step change in each of the parameters of the gamma-Poisson mixture model following a genuine control chart signal. After a control chart detects the process change, the engineer can very quickly isolate when the true change may have occurred by using the estimated change point as a guide.

1.3 Research Objectives and Assumptions

The objectives of this research are as follows:

1. Derive the gamma-Poisson mixture model that will be used throughout this research effort.
2. Derive a change-point estimator for the time of step change in each of the parameters of the gamma-Poisson mixture model using the method of maximum likelihood estimation.
3. Use Monte Carlo simulation to evaluate the root mean square error performance of the change-point estimators.
4. Present and evaluate an approach based on the likelihood function for estimating a confidence set for the process change point.

The assumptions made in this research are as follows:

1. This research effort will investigate change-point estimation techniques for the two-parameter gamma-Poisson mixture model. In the first case, it is assumed that the process experiences a step change in the location parameter following some unknown change point τ. That is, the process starts in an in-control state where the actual in-control mean is known or sufficiently estimated from a phase I study. After some unknown point in time, τ, the process mean suddenly shifts to some unknown, out-of-control value and remains at that value until detected and corrected by the process engineer.
2. For the second case, the assumption is that the process experiences a step change in the overdispersion parameter following some unknown change point τ. That is, the process starts in an in-control state where the actual in-control overdispersion parameter is known or sufficiently estimated from a phase I study. After some unknown point in time, τ, the overdispersion parameter suddenly shifts to some unknown, out-of-control value and remains at that value until detected and corrected by the process engineer.

1.4 Thesis Organization

This thesis is divided into five chapters. Chapter 1 presented relevant background information on SPC, the motivation for this research, and the problem to be studied. Then, the research objectives and assumptions of this thesis were presented. Chapter 2 reviews relevant literature concerning change-point estimation within an SPC context. Chapter 3 develops the mathematical foundation for the proposed change-point

estimators. Chapter 4 discusses the Monte Carlo simulation study that was used to evaluate the performance of the proposed methodologies, and presents the results. Last, Chapter 5 presents the conclusions of this research and proposes recommendations for future research.

2. Literature Review

Pignatiello [16] points out that very little attention has been given to change-point estimation in the quality engineering literature. This chapter presents relevant change-point estimation research, as applied within an SPC context, which has been given attention in the literature. This literature shows processes that experience step changes, linear trends, and monotonic changes have been studied for normal, Poisson, and/or binomial processes. First, this chapter starts with a review of a change-point model designed for step changes.

2.1 Overview of a Change-Point Model

The change-point model used in this research is attributed to Hinkley [7]. Hinkley proposed a change-point model for step changes in normal process means. Montgomery [9] states that the change-point procedure is very good at estimating the point in time when a process has changed, and that it should be given wider attention in process monitoring problems. Montgomery gives an introduction to the change-point model, under normal theory assumptions, applied within an SPC context. Now a change-point model for step changes is defined.

Consider a process where a sustained shift has occurred in a parameter, Θ. The change-point model is defined as

$$\begin{aligned} y_i &\sim f_0(\Theta_0),\ \mathrm{i} = 1, 2, ..., \tau \\ y_i &\sim f_a(\Theta_a),\ \mathrm{i} = t+1, t+2, ...T \end{aligned} \tag{2.1}$$

where the in-control distribution is $f_0(\Theta_0)$ up to and including sample τ, the unknown change point. The sustained parameter shift occurs between τ and $\tau+1$, and

observations are sampled from the out-of-control distribution $f_a(\Theta_a)$ from time $\tau+1$ to *T*. Note, Θ can represent a vector of parameters, as well as a single parameter. Now that the basic change-point model has been introduced, some relevant literature on change-point estimation, within an SPC context, will be discussed.

2.2 Change-Point Estimation

This section will review relevant literature on change-point estimation considered in the quality engineering literature. The reader is referred to Basseville and Nikiforov [1] for a more general coverage of change-point detection and estimation algorithms.

Three probability models often used in SPC applications are the normal, Poisson, and binomial distributions. The normal distribution is probably the most important distribution in the theory and application of statistics. For instance, normal theory is the foundation for Shewhart $\bar{x}$ charts. The Poisson probability model is an important discrete distribution used in SPC. In a manufacturing context, for example, the Poisson model is often used to model the number of defects (or nonconformities) that occur within an inspection unit. Likewise, the binomial probability model is another important discrete distribution used in SPC. This discrete distribution is often an appropriate model for sampling from a large population, where parameter *p* represents the fraction of defective (or nonconforming) items in the population. The relevant change-point estimation literature concerning these distributions is discussed next.

2.2.1 Normal Case

Samuel, Pignatiello, and Calvin [21] consider monitoring a process with a Shewhart $\bar{x}$ control chart. They suggest using the method of maximum likelihood estimation to estimate the time of a step change in the mean of the process. These authors acknowledge that a Shewhart control chart can signal some time later from when a change actually occurred in the process mean. Thus, they argue the importance of having an estimate of the process change point. They derive their estimator and then evaluate its performance using Monte Carlo simulation. They show that applying the MLE following a genuine Shewhart control chart signal provides an accurate estimate for the change point.

Khoo [8] used the MLE proposed by Samuel, Pignatiello, and Calvin [21] to identify the time of a permanent shift in the mean of a process being monitored by CUSUM control charts. He concluded that using this change-point estimator with the CUSUM, rather than the Shewhart $\bar{x}$ control chart, offers some advantages. Namely, for small shifts, the expected run length is lower than that of the $\bar{x}$ chart, therefore allowing process improvement to be carried out earlier.

Pignatiello and Samuel [16] compare the MLE of the time of a step change in a normal process mean to the built-in change-point estimators offered by the CUSUM and exponentially weighted moving average (EWMA) control charts. Nishina [10] points out that the CUSUM, EWMA, and moving average charts are similar with respect to estimating a process change point, but that the CUSUM is more efficient than the other two in change-point estimation. Using Monte Carlo simulation, Pignatiello and Samuel

[16] show that the performance of the MLE is better than the built-in estimators of both the EWMA and CUSUM when the entire range of step change magnitudes is considered. Not only do they consider just the point estimate, but they also propose a confidence set for the time of the process change. The estimated change point denotes the most likely location for a step change in the parameter of interest, while the confidence set provides a window of potential locations to search for an assignable cause. A confidence set on the process change point is an advantage of using the MLE as opposed to the CUSUM or EWMA change point estimators. Note, while EWMA control charts were not considered for this thesis, an interested reader may refer to Roberts [18] or Montgomery [9] for further details on this control chart.

Rather than assuming the out-of-control process behavior is adequately modeled by a step function, Perry and Pignatiello [12] derive the MLE for the time of linear trend change in a normal process mean. They compared performance results between their estimator and that suggested by Samuel, Pignatiello, and Calvin [21] following genuine $\bar{x}$ control chart signals when a linear trend disturbance was present. Perry and Pignatiello [12] use Monte Carlo simulation to evaluate the performance of their estimator. They show that for a process where a linear trend is present, the MLE of the process change point derived for linear trends outperforms the MLE for step changes. They also evaluated the performance of the proposed estimator when the distribution of the random variable being monitored is not exactly normal. It was shown that the estimator performs well when the distribution is symmetric, but heavier-tailed than the normal. Last, they presented an approach using the likelihood function for estimating a

confidence set. This would provide process engineers a smaller subset of observations to investigate for an assignable cause of the process change.

Park and Park [11] argue that the $\bar{x}$ chart is affected by changes in the process variance as well as the process mean. For example, many of the authors cited above will estimate the time of process change based on a change in the mean alone. Park and Park state that when the $\bar{x}$ chart issues a signal, a wrong conclusion could be reached if the change-point estimator considers a change in the process mean only. Thus, they apply a joint estimator of a change point to the $\bar{x}$ and S control charts, which considers the change in the process mean and variance simultaneously. Using Monte Carlo simulation they showed that the proposed estimator performed well irrespective of subgroup size over various changes in both the process mean and variance.

2.2.2 Poisson Case

Samuel and Pignatiello [20] derived their estimator for a Poisson process under the assumption of a step change in the Poisson rate parameter λ. They applied the proposed estimator after a signal from a *c*-chart. Using Monte Carlo simulation to evaluate the performance of their proposed estimator, they concluded overall good accuracy and precision performance. They demonstrated that the estimator improves as the magnitude of the step change increases.

Perry, Pignatiello, and Simpson [15] derived and evaluated the MLE of the process change point using the change likelihood function for a linear trend in a Poisson rate. As expected, their proposed estimator provides good overall accuracy and precision performances for a process that experiences a linear trend relative to the MLE derived for

step changes. They note that if the MLE derived for step changes is applied to processes that experience a linear trend, the estimator will tend to overestimate the true change point.

As discussed above, many different change-point methods have been suggested. Most of these methods assume the type of change in a process parameter is known before hand. Perry, Pignatiello, and Simpson [14] propose a maximum likelihood estimator for the time of change of a Poisson rate parameter without requiring exact knowledge, *a priori*, regarding the exact form of the change present. They evaluated the performance of their estimator relative to that achieved by the estimator suggested by Samuel and Pignatiello [20] following genuine Poisson CUSUM control chart signals and across several types of monotonic behavior in the rate parameter. These monotonic functions include sudden step, linear trend, exponential, and log functions. Perry, Pignatiello, and Simpson [14] concluded it is better to use their proposed estimator when the form of the change present is only known to be monotonic. Their proposed method is important since process engineers rarely know the type of change before hand.

2.2.3 Binomial Case

The discussions above on change-point estimation have focused on normal and Poisson processes that experience step changes, linear trends, and even the more general monotonic changes. Regarding change-point estimation for binomial processes, Pignatiello and Samuel [17] consider a step change in p, the process fraction nonconforming. They propose an estimator using the method of maximum likelihood estimation to estimate when the change in p began. This estimator can be applied after a

p-chart or an *np*-chart has detected a change. Monte Carlo simulation was used to show that the estimator provides accurate and precise estimates of the time of a step change when the magnitude of the change in p is at least 10% and larger. They also stress the ease of using this method in a spreadsheet application.

Perry and Pignatiello [13] also considered a process that experiences a step change in the process fraction nonconforming. They compared the performance of the MLE proposed by Pignatiello and Samuel [17] to built-in change-point estimators from binomial CUSUM and EWMA control charts. They concluded that the MLE of the change point will provide an overall better estimate of the process change point when compared to the built-in estimates of the CUSUM and EWMA. They do point out that the only exception to this is when the process changes to the pre-specified magnitude of change used to design the CUSUM or EWMA control charts. However, since these magnitudes are rarely known before hand, the MLE is still the better choice overall.

2.3 Conclusion

This chapter started with a review of a change-point model for step changes. Next this chapter reviewed some relevant literature on change-point estimation as applied within an SPC context. As discussed, processes that experience step changes, linear trends, and monotonic changes in general have been studied for normal, Poisson, and/or binomial processes. However, the case of estimating a change point in a process parameter, where the observations are overdispersed counts, has currently not received much, if any, attention in the quality engineering literature. Hence, that is the focus of this research. The methodology of this research will be developed next.

3. Methodology

This chapter focuses on the mathematical models used for this research. First, the gamma-Poisson mixture model is derived. It is the probability model that will be the basis for modeling overdispersed count data in this thesis. Then, two different step-change scenarios will be addressed. The change-point model for a process that experiences a step change in mean is defined, followed by the derivation of the MLE of the process change point. Next, the change-point model for a process that experiences a step change in the overdispersion parameter of the underlying probability distribution is defined. Then, the corresponding MLE of the process change point for this parameter will be derived. Last, the method of building a confidence set to capture the true change point is discussed.

3.1 Derivation of the gamma-Poisson Mixture Model

The Poisson distribution was defined previously in equation (1.4) to introduce the idea of overdispersion. For completeness of deriving the gamma-Poisson mixture model, it will be restated here.

Let Y_i denote a random variable distributed as Poisson(λ_i), where $\lambda_i > 0$ is the rate parameter at time *i* and denotes the expected value of Y_i. The Poisson probability mass function (PMF) is a discrete function defined for $y_i \in [0,\ 1,\ 2,\ ...]$ given by

$$P(Y_i = y_i) = f(y_i \mid \lambda_i) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} \tag{3.1}$$

where, $P(Y_i = y_i \mid \lambda_i)$ denotes the probability that the random variable Y_i will take on the value y_i, given λ_i.

Suppose that it is postulated that $\lambda_i = \lambda_0$ for all $i = 1, 2, ..., T$. Then under Poisson model assumptions, one should expect that $E(Y_i) = Var(Y_i) = \lambda_0$, i.e., *equidispersion*, for all $i = 1, 2, ..., T$. Unfortunately, this is rarely the case in practice. Instead it is often true that the observed variance of a random sample is greater than the theoretical variance, implying that the count data are overdispersed. If Poisson count data are collected over time in discrete samples, overdispersion can occur as a result of uncontrollable fluctuations in the rate parameter over this time domain.

It is reasonable to assume the variability in λ_i is well modeled using the flexible, two-parameter gamma probability distribution. Let $\lambda_i > 0$ be distributed gamma(α, β), where $\beta = \frac{\nu}{\alpha}$, $\alpha > 0$ is the shape parameter, and $\beta > 0$ is the scale parameter. The gamma PDF is a continuous distribution with non-negative support defined as

$$f(\lambda_i \mid \alpha, \beta) = \frac{\lambda_i^{\alpha-1} e^{\frac{-\lambda_i}{\beta}}}{\Gamma(\alpha)\beta^{\alpha}}. \tag{3.2}$$

If λ_i denotes a random variable possessing a gamma distribution with parameters α and β for $i = 1, 2, ..., T$, then it is easily shown that $E(\lambda_i) = \nu$ and $Var(\lambda_i) = \frac{\nu^2}{\alpha}$. In equation (3.2), $\Gamma(z)$ denotes the gamma function where for a non-negative integer z, $\Gamma(z+1) = z!$, where ! denotes the factorial function.

The joint density function, using the parameterization $\beta = \frac{\nu}{\alpha}$, is given as

$$f(y_i, \lambda_i) = f(y_i \mid \lambda_i) f(\lambda_i)$$

$$= \frac{e^{-\lambda_i} \lambda_i^{(y_i+\alpha-1)} e^{-\frac{\alpha\lambda_i}{\nu}} \left(\frac{\nu}{\alpha}\right)^{-\alpha}}{\Gamma(y_i+1)\Gamma(\alpha)}. \tag{3.3}$$

Thus, the marginal distribution of y_i for $i = 1, 2, ..., T$ is

$$f(y_i) = \int_0^\infty f(y_i, \lambda_i) d\lambda_i$$

$$= \int_0^\infty \frac{e^{-\lambda_i} \lambda_i^{(y_i+\alpha-1)} e^{\frac{-\alpha\lambda_i}{\nu}} \left(\frac{\nu}{\alpha}\right)^{-\alpha}}{\Gamma(y_i+1)\Gamma(\alpha)} d\lambda_i$$

$$= \frac{\Gamma(1+\frac{\alpha}{\nu})^{(-y_i-\alpha)} (\frac{\nu}{\alpha})^{-\alpha} \Gamma(y_i+\alpha)}{\Gamma(y_i+1)\Gamma(\alpha)}$$

$$= \frac{\Gamma(y_i+\alpha)}{\Gamma(y_i+1)\Gamma(\alpha)} \left(\frac{\nu}{\alpha+\nu}\right)^{y_i} \left(\frac{\alpha}{\alpha+\nu}\right)^{\alpha} \tag{3.4}$$

which is the well-known gamma-Poisson mixture parameterization of the negative binomial distribution with parameters $r > 0$, where $r = \alpha$ and $p \in [0,1]$, where $p = \frac{\alpha}{\alpha+\nu}$. This model is employed as the underlying model for the data in this research. The negative binomial model in equation (3.4) can also be written in the more familiar form as

$$K_i (1-p)^{y_i} p^r \tag{3.5}$$

where the constant $K_i = \frac{\Gamma(y_i+r)}{\Gamma(y_i+1)\Gamma(r)}$, $(1-p) = \frac{\nu}{r+\nu}$, and $p = \frac{r}{r+\nu}$. For the remainder of this thesis, the expanded form shown in equation (3.4) will be used. Note, by rewriting equation (3.4) as

$$f(y_i) = \frac{\Gamma(y_i + \alpha)}{\Gamma(\alpha)(\alpha + \nu)^{y_i}} \left(\frac{1}{1 + \frac{\nu}{\alpha}} \right)^{-\alpha} \left(\frac{\nu^{y_i}}{y_i!} \right)$$

and calculating $\lim_{\alpha \to \infty} f(y_i)$ yields

$$\lim_{\alpha \to \infty} f(y_i) = \frac{e^{-\nu} \nu^{y_i}}{y_i!}$$

which is the Poisson probability mass function. This shows that the Poisson distribution is a limiting form of the negative binomial distribution as α goes to infinity. Therefore, α can be interpreted as a parameter accounting for any overdispersion that may have occurred in the data. Hence, α is referred to as the overdispersion parameter. Next, the mean and variance of the negative binomial distribution are calculated from the moment generating function.

The moment generating function of $f(y_i)$ is

$$m_y(t) = E(e^{ty}) = \sum_y e^{ty} f(y_i)$$

$$= \sum_{y_i} e^{ty_i} \frac{\Gamma(y_i + \alpha)}{\Gamma(y_i + 1)\Gamma(\alpha)} \left(\frac{\nu}{\alpha + \nu} \right)^{y_i} \left(\frac{\alpha}{\alpha + \nu} \right)^{\alpha}$$

$$= \left(\frac{\alpha}{\alpha + \nu} \right)^{\alpha} \left(\frac{\alpha + \nu - e^t \nu}{\alpha + \nu} \right)^{(-\alpha)}$$

$$= \left(\frac{\alpha}{\alpha + \nu - e^t \nu} \right)^{\alpha}. \tag{3.6}$$

The first moment of Y_i about the origin (i.e., the expected value or mean) is calculated as

$$u_1^{'} = E(Y_i) = m^{(1)}(0)$$

$$= \nu. \tag{3.7}$$

Likewise, the second moment about the origin of Y_i is

$$u_2^{'} = E(Y_i^2) = m^{(2)}(0)$$

$$= \nu^2 + \tfrac{\nu^2}{\alpha} + \nu. \tag{3.8}$$

Thus, the variance of Y_i can be calculated as

$$Var(Y_i) = E(Y_i^2) - E(Y_i)^2$$

$$= (\nu^2 + \tfrac{\nu^2}{\alpha} + \nu) - \nu^2$$

$$= \tfrac{\nu^2}{\alpha} + \nu. \tag{3.9}$$

Note that if $\lim_{\alpha \to \infty}(\tfrac{\nu^2}{\alpha} + \nu)$ is calculated, $Var(Y_i) = \nu = E(Y_i)$. Thus, as α goes to infinity, the variance approaches the theoretical variance of a Poisson random variable. Now that the probability model for this research has been derived, the change-point models can be defined and the proposed estimators derived.

3.2 Change-Point Model for a Step Change in Mean

The change-point model for a step change in mean is parameterized as follows:

$$\begin{aligned} y_i &\sim NB(\alpha, \tfrac{\alpha}{\alpha+\nu_0}),\ \mathrm{i} = 1, 2, ..., \tau \\ y_i &\sim NB(\alpha, \tfrac{\alpha}{\alpha+\nu_a}),\ \mathrm{i} = \tau+1, \tau+2, ...T. \end{aligned} \tag{3.10}$$

This model assumes an in-control process with distribution NB($\alpha, \tfrac{\alpha}{\alpha+\nu_0}$) up to and including a point τ, the unknown change point. Assume α, the overdispersion parameter, and ν_0, the in-control mean, are both known or sufficiently estimated. Between times τ and $\tau+1$ the process shifts to a new level, $\nu_a \neq \nu_0$, and the process

continues with a sustained shift in mean with the distribution NB($\alpha, \frac{\alpha}{\alpha+\nu_a}$). For this research, ν_a is the unknown mean after the process experiences the step change, and *T* denotes the time of the control chart signal. This change-point model can be used to derive a maximum likelihood estimator for the process change point, which is shown next.

3.3 Derivation of the MLE When ν_a is Unknown

The proposed MLE of the change point is denoted as $\hat{\tau}_\nu$. Assuming a process change point at τ, the likelihood function, derived from equation (3.4) is

$$L(\tau, \nu_a \mid y_i) = \prod_{i=1}^{\tau} K_i \left(\frac{\nu_0}{\alpha+\nu_0}\right)^{y_i} \left(\frac{\alpha}{\alpha+\nu_0}\right)^{\alpha} \prod_{i=\tau+1}^{T} K_i \left(\frac{\nu_a}{\alpha+\nu_a}\right)^{y_i} \left(\frac{\alpha}{\alpha+\nu_a}\right)^{\alpha} \tag{3.11}$$

where $K_i = \frac{\Gamma(y_i+\alpha)}{\Gamma(y_i+1)\Gamma(\alpha)}$.

Assuming there has been a single change point, the goal is to estimate τ, the unknown change point. The MLE of τ is the value of τ that maximizes the likelihood function in equation (3.11), or equivalently its logarithm. Taking the natural logarithm of equation (3.11) and simplifying (less the constant K_i) yields the log-likelihood function

$$l(\tau, \nu_a \mid y_i) = \tau\left[\bar{y}_\tau \ln\left(\frac{\nu_0}{\alpha+\nu_0}\right) + \alpha \ln\left(\frac{\alpha}{\alpha+\nu_0}\right)\right] + (T-\tau)\left[\bar{y}_{T,\tau} \ln\left(\frac{\nu_a}{\alpha+\nu_a}\right) + \alpha \ln\left(\frac{\alpha}{\alpha+\nu_a}\right)\right]. \tag{3.12}$$

Again, the objective is to estimate the unknown change point τ; however, there are two unknowns in the log-likelihood function: τ and ν_a. An estimate for ν_a is needed to maximize the log-likelihood function. If the value of τ was known, then the value of ν_a that maximizes the log-likelihood function is given by the solution to

$$\frac{\partial l}{\partial \nu_a} = \frac{\bar{y}_{T,\tau}(T-\tau)\left[\frac{1}{\alpha+\nu_a} - \frac{\nu_a}{(\alpha+\nu_a)^2}\right](\alpha+\nu_a)}{\nu_a} - \frac{(T-\tau)\alpha}{\alpha+\nu_a} = 0.$$

Solving this in terms of ν_a yields

$$\hat{\nu}_a(\tau) = \bar{y}_{T,\tau}$$

where $\bar{y}_{T,\tau} = (T-\tau)^{-1}\sum_{i=\tau+1}^{T} y_i$. Making this substitution into the log-likelihood function of equation (3.12) gives

$$l(\tau \mid y_i) = \tau\left[\bar{y}_\tau \ln(\tfrac{\nu_0}{\alpha+\nu_0}) + \alpha \ln(\tfrac{\alpha}{\alpha+\nu_0})\right] + (T-\tau)\left[\bar{y}_{T,\tau} \ln(\tfrac{\bar{y}_{T,\tau}}{\alpha+\bar{y}_{T,\tau}}) + \alpha \ln(\tfrac{\alpha}{\alpha+\bar{y}_{T,\tau}})\right]. \qquad (3.13)$$

Now the log-likelihood function has only one unknown parameter, the change point τ. Evaluating equation (3.13) over all possible integer change points $t \in [0,1,...T-1]$ in search of the maximum yields

$$\hat{\tau}_\nu = \arg\max_{0 \le t < T} \{l(t \mid y_i)\} \qquad (3.14)$$

where $\hat{\tau}_\nu$ is the maximum likelihood estimate of the last observation taken from the in-control process. The MLE of τ will be applied to data obtained following a control chart signal. To make this concept more concrete, consider the simulated example shown next.

3.3.1 MLE Example

Consider a fabric manufacturing process. The quality characteristic being measured is the number of defects in each 6'x10' section of fabric produced. The manufacturing process is very difficult to get to the point where there are no defects. In fact the in-control mean number of defects is two per 60 square feet of fabric, i.e. $\nu_0 = 2$. For this example, the overdispersion parameter, α, is assumed to be 10. Calculating the control limits for the 3σ control chart yields UCL = 6.64 and LCL = 0. If the number of defects equals 7 or greater, then the fabric will be considered defective and will be scrapped. The number of defects per unit of fabric is shown in Table 1.

Table 1: Observed Count at Each Observation i

Observation i	Number of Defects	Observation i	Number of Defects
1	2	15	0
2	1	16	3
3	2	17	3
4	2	18	2
5	3	19	1
6	1	20	0
7	0	21	1
8	2	22	1
9	1	23	2
10	4	24	1
11	0	25	1
12	3	26	3
13	3	27	4
14	0	28	7

Twenty seven units of fabric were inspected and determined to be within the control limits and the 28^{th} unit of fabric exceeded the UCL. Thus, the 3σ control chart signaled at time T = 28. The proposed estimator was applied after the control chart

signaled. In applying the proposed change-point estimator, each potential change point, *t*, needs to be used to calculate its corresponding log-likelihood value from equation (3.13).

To evaluate equation (3.13), first all the averages, $\bar{y}_{T,t} = (T-t)^{-1}\sum_{i=t+1}^{T} y_i$, and $\bar{y}_t = (t)^{-1}\sum_{i=1}^{t} y_i$ need to be calculated. Recall in this example that *T* = 28. To evaluate equation (3.13) for, say, potential change point *t* = 26, calculate $\bar{y}_{T,t}$, and $\bar{y}_t$ as $\bar{y}_{28,26} = (2)^{-1}\sum_{i=27}^{28} y_i = \frac{4+7}{2} = 5.5$ and $\bar{y}_{26} = (26)^{-1}\sum_{i=1}^{26} y_i = \frac{2+1+...+1+3}{26} = 1.6154$.

Bringing these values and the other known parameters into equation (3.13) yields a log-likelihood value of -142.82 for *t* = 26. Evaluating equation (3.13) over all *t* will provide the value of *t* that achieves the maximum log-likelihood value, which is $\hat{\tau}_v = 26$. All log-likelihood values for this example are shown in Table 2.

In this example *t* = 26 gives the largest log-likelihood value. That is, the proposed estimator calculates the change point at $\hat{\tau}_v = 26$. The conclusion is that observation *i* = 26 is the last observation from the in-control process and observation *i* = 27 is the first observation from the out-of-control process. Process engineers would be instructed to look for a change in the process between observations 26 and 27. Note that the simulated change point for this example is τ = 25. In this example, the point estimate is off by one, which also alludes to the usefulness of confidence sets (covered later) to capture the true change point. A similar example was done by Samuel, Pignatiello, and Calvin [21] for a step change in the normal process mean.

Table 2: Log-Likelihood Values at each Potential Change Point

Observation i	Number of Defects	t	Log - Likelihood Value
1	2	0	-145.9449
2	1	1	-145.9423
3	2	2	-145.9807
4	2	3	-145.9794
5	3	4	-145.9780
6	1	5	-145.9296
7	0	6	-145.9747
8	2	7	-146.0133
9	1	8	-146.0133
10	4	9	-146.0024
11	0	10	-146.0016
12	3	11	-146.0012
13	3	12	-146.0133
14	0	13	-145.9992
15	0	14	-145.9986
16	3	15	-145.8752
17	3	16	-145.9460
18	2	17	-145.9947
19	1	18	-145.9928
20	0	19	-145.9245
21	1	20	-145.6326
22	1	21	-145.3570
23	2	22	-144.9582
24	1	23	-144.7837
25	1	24	-144.0714
26	3	25	-142.9806
27	4	26	**-142.8196**
28	7	27	-143.1652

3.4 Change-Point Model for a Step Change in the Overdispersion Parameter

The previous section defined the change-point model and derived the corresponding estimator for the case where a step change occurred in the mean of the process. This section considers the case where the process experiences a step change in the overdispersion parameter, α. The change-point model is parameterized as follows:

$$y_i \sim NB(\alpha_0, \tfrac{\alpha_0}{\alpha_0+\nu}),\ \mathrm{i} = 1,2,...,\tau$$
$$y_i \sim NB(\alpha_a, \tfrac{\alpha_a}{\alpha_a+\nu}),\ \mathrm{i} = \tau+1,\tau+2,...,T. \qquad (3.15)$$

This model assumes an in-control process with distribution NB($\alpha_0, \frac{\alpha_0}{\alpha_0+\nu}$) up to and including the unknown change point, τ. Assume α_0, the in-control overdispersion parameter, and ν, the process mean, are both known. Between times τ and $\tau+1$ the process experiences a shift in the overdispersion parameter to a new level, $\alpha_a \neq \alpha_0$, and the process continues with a sustained shift with the distribution NB($\alpha_a, \frac{\alpha_a}{\alpha_a+\nu}$). For this research, α_a is the unknown overdispersion parameter after the process experiences the step change, and *T* denotes the time of the control chart signal. This change-point model is used to derive a maximum likelihood estimator for the process change point.

3.5 Derivation of the MLE When α_a is Unknown

Assuming a process change point at τ, the likelihood function derived from equation (3.4) is

$$L(\tau, \alpha_a \mid y_i) = \prod_{i=1}^{\tau} \frac{\Gamma(y_i+\alpha_0)}{\Gamma(y_i+1)\Gamma(\alpha_0)} \left(\frac{\nu}{\alpha_0+\nu}\right)^{y_i} \left(\frac{\alpha_0}{\alpha_0+\nu}\right)^{\alpha_0} \prod_{i=\tau+1}^{T} \frac{\Gamma(y_i+\alpha_a)}{\Gamma(y_i+1)\Gamma(\alpha_a)} \left(\frac{\nu}{\alpha_a+\nu}\right)^{y_i} \left(\frac{\alpha_a}{\alpha_a+\nu}\right)^{\alpha_a}. \qquad (3.16)$$

The goal is to estimate τ, under the assumption that there is a single, unknown change point. The MLE of τ is the value of τ that maximizes the likelihood function in equation (3.16), or equivalently its logarithm. Taking the natural logarithm of equation (3.16) and simplifying (less a constant, $\ln \frac{\nu^{y_i}}{\Gamma(y_i)}$) yields the log-likelihood function

$$l(\tau,\alpha_a \mid y_i)=\sum_{i=1}^{\tau}\left[\ln\left(\frac{\Gamma(y_i+\alpha_0)}{\Gamma(\alpha_0)}\right)-\ln(\alpha_0+\nu)(y_i+\alpha_0)+\alpha_0\ln(\alpha_0)\right]+ \sum_{i=\tau+1}^{T}\left[\ln\left(\frac{\Gamma(y_i+\alpha_a)}{\Gamma(\alpha_a)}\right)-\ln(\alpha_a+\nu)(y_i+\alpha_a)+\alpha_a\ln(\alpha_a)\right]. \tag{3.17}$$

Again, the objective is to estimate the unknown change point, τ; however, there are two unknowns in the log-likelihood function of equation (3.17): τ and α_a. A value for α_a is needed to maximize this log-likelihood function. If the value of τ were known, then the value of α_a that maximizes the log-likelihood function is given by the solution to

$$\frac{\partial l}{\partial \alpha_a}=\sum_{i=\tau+1}^{T}\left[\psi(y_i+\alpha_a)-\psi(\alpha_a)-\frac{(y_i+\alpha_a)}{(\alpha_a+\nu)}+\ln\left(\frac{\alpha_a}{\alpha_a+\nu}\right)+1\right]=0 \tag{3.18}$$

where $\psi(x)=\frac{\partial \ln\Gamma(x)}{\partial x}$.

Since there is no closed-form solution for α_a, Newton's method will be used to evaluate equation (3.18) over each potential change-point value. This provides an estimate of α_a for each potential τ. Substituting an estimate, $\hat{\alpha}_a(\tau)$, for each α_a in equation (3.17) yields

$$l(\tau \mid y_i)=\sum_{i=1}^{\tau}\left[\ln\left(\frac{\Gamma(y_i+\alpha_0)}{\Gamma(\alpha_0)}\right)-\ln(\alpha_0+\nu)(y_i+\alpha_0)+\alpha_0\ln(\alpha_0)\right]+ \sum_{i=\tau+1}^{T}\left[\ln\left(\frac{\Gamma(y_i+\hat{\alpha}_a)}{\Gamma(\hat{\alpha}_a)}\right)-\ln(\hat{\alpha}_a(\tau)+\nu)(y_i+\hat{\alpha}_a(\tau))+\hat{\alpha}_a(\tau)\ln(\hat{\alpha}_a(\tau))\right]. \tag{3.19}$$

Now the log-likelihood function in equation (3.19) has only one unknown parameter, the change point, τ. Evaluating equation (3.19) over all possible integer change points $t\in[0,1,...,T-1]$ in search of the maximum yields

$$\hat{\tau}_\alpha=\arg\max_{0\le t<T}\{l(t \mid y_i)\} \tag{3.20}$$

where $\hat{\tau}_\alpha$ is the maximum likelihood estimate of the last observation taken from the in-control process. The next section gives a short summary of Newton's method and how it was used to estimate α_a in this thesis.

3.6 Newton's Method for Finding the MLE of α_a

Newton's method is a derivative-based algorithm that uses a linear approximation to find roots (zeros) for real-valued functions. Consider finding the root of some differentiable function f. Given some current approximation, x_k, a better approximation, x_{k+1}, may be found using

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}. \tag{3.21}$$

Given an initial guess, x_0, chosen near the root of the function f, and an appropriate stopping scheme, equation (3.21) should converge at the root of *f*. The reader is referred to Burden and Faires [3] for more detail regarding Newton's method. Now, some discussion is provided on the application of Newton's method to the problem presented in this thesis.

If τ was known, Newton's method could be used to solve for α_a in equation (3.18). The estimate, $\hat{\alpha}_a$, at the k+1st iteration can be calculated explicitly from

$$\hat{\alpha}_{a,k+1} = \hat{\alpha}_{a,k} - \frac{\sum_{i=\tau+1}^{T}\left[\psi(y_i+\hat{\alpha}_{a,k}) - \psi(\hat{\alpha}_{a,k}) - \frac{(y_i+\hat{\alpha}_{a,k})}{(\hat{\alpha}_{a,k}+\nu)} + \ln\left(\frac{\hat{\alpha}_{a,k}}{\hat{\alpha}_{a,k}+\nu}\right)+1\right]}{\sum_{i=\tau+1}^{T}\left[\psi(1, y_i+\hat{\alpha}_{a,k}) - \psi(1,\hat{\alpha}_{a,k}) + \frac{\hat{\alpha}_{a,k}y_i+\nu^2}{\hat{\alpha}_{a,k}(\hat{\alpha}_{a,k}+\nu)^2}\right]} \tag{3.22}$$

where $\alpha_{a,0}$ is the initial guess, $\psi(x) = \text{digamma function} = \frac{\partial \ln \Gamma(x)}{\partial x}$, and $\psi(1,x)$ is the trigamma function. For this thesis, the initial guess was chosen to be $\alpha_{a,0} = 0.01$.

Estimates for α_a can be obtained at each potential change point, $t \in [0,1,...T-1]$, using the iterative procedure defined in equation (3.22) given the vector of observations y. The stopping scheme should terminate the iterative procedure if $\hat{\alpha}_{a,k} - \hat{\alpha}_{a,k+1}$ is sufficiently small. For example, if the square root of the sum of squares of the difference of the last two iterations (i.e., the 2-norm) is less than, say, 1.0 x 10^{-6}. The estimator from equation (3.22) is then used for the evaluation of the argument of equation (3.20). An example of Newton's method, coded in MATLAB, is provided in Figure 4.

```
a_k = 0.01; %Initial guess
tol = 10^(-6); %Tolerance
N = 80; %Number of iterations

     for k = 1:N
       %Numerator of equation (3.22)
       P=0;
       P = sum((psi(0,y(t+1:T)+a_k) - psi(0,a_k) - ...
           ((y(t+1:T)+a_k)./(a_k + v)) + log(a_k/(a_k+v)) + 1));

        %Denominator of equation (3.22)
       Q=0;
       Q = sum((psi(1, y(t+1:T)+a_k) - psi(1, a_k) + ...
           ((a_k.*y(t+1:T) + v^2)/(((a_k+v)^2)*a_k))));

        %Newton steps
       a_k1 = a_k + (-P/Q);
       error = norm(a_k - a_k1, 2);
       a_k = a_k1;

       if error < tol
          break
       End
       if a_k <= 0 %Constraint to keep a_k positive
          a_k = 10,000; %Set "large"
          break
       end
    end
```

Figure 4: Newton's Method Coded in MATLAB

3.6.1 Additional Newton Method Constraint

It was noticed in some instances of Newton's method that $\hat{\alpha}_a$ grows very large and then turns negative due to a sign change in the numerator of equation (3.22) (denoted as P in Figure 4). However, it is defined that $\alpha > 0$. Therefore, the additional constraint, shown in the code of Figure 4, was added to Newton's method to keep the estimates of α_a positive. Since the maximum of the log-likelihood function of (3.17) in these cases is achieved for very large α_a, $\hat{\alpha}_a$ is set to a large value (i.e., $\hat{\alpha}_a = 10,000$). Figure 5 shows the effect of increasing α over the log-likelihood function, as well as the numerator (P) and denominator (Q) of equation (3.22) for a single observation *y*. Notice that the denominator (Q) of (3.22) is always negative, which is a necessary condition for (3.17) to have a maximum.

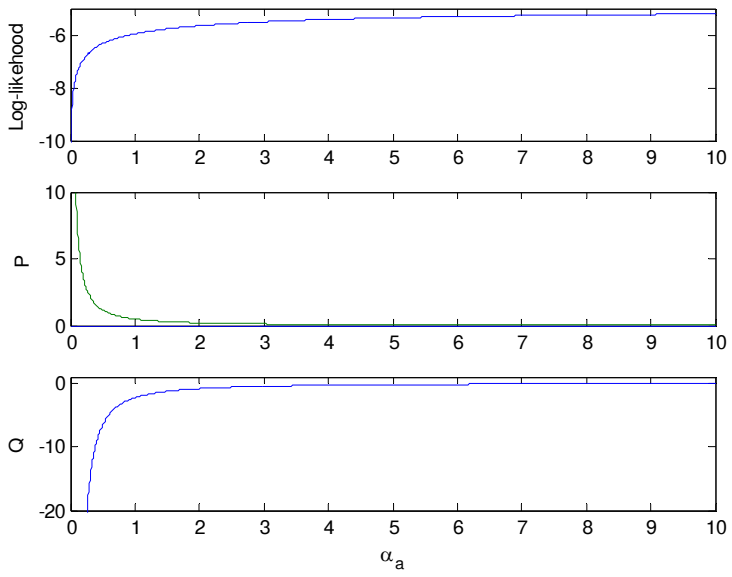

Figure 5: Plots Showing the Effect of Increasing α Over the Log-likelihood Function, and P and Q from MATLAB code in Figure 4 for Single Observation y. ν=5 and y=5.

3.7 Confidence Sets

Estimating $\hat{\tau}$ only gives a point estimate. This research now constructs confidence sets on the process change points, which can provide the process engineer a search window of potential change points that covers the true process change point with a given level of confidence. The method of constructing confidence regions on parameter estimates using the likelihood function is attributed to Box and Cox [2]. The confidence set is defined as follows:

$$CS = \{t : \ln L(t) > \ln L(\hat{\tau}) - D\} \tag{3.23}$$

where $\ln L(\hat{\tau})$ is the maximum of the log-likelihood function evaluated over all potential change points *t*. If at a given *t*, the value of $\ln L(t)$ is greater than $\ln L(\hat{\tau}) - D$, where *D* is some specified constant, then *t* is included in the confidence set. Box and Cox suggest using a reference value $D = \frac{1}{2}\chi^2_{\alpha,1}$ to obtain a $100(1-\alpha)\%$ confidence region. However, surface plots are provided in Chapter 4 to aid the user in selecting a value for *D* that will more accurately achieve the user's desired level of confidence. This confidence set approach is the same approach used by Pignatiello and Samuel [16].

3.7.1 Confidence Set Example

To make equation (3.23) more concrete, consider the following example. A negative binomial process has in-control parameters NB($\alpha, \frac{\alpha}{\alpha+\nu_0}$), where $\alpha = 10$, and $\nu_0 = 10$. Following the 50th observation, the process experiences a step change in mean where $\nu_a = 14$. In this example, $\hat{\tau}_\nu$, as estimated from equation (3.14), yields $\hat{\tau}_\nu$ = 47,

whereas the true change point is actually, $\tau = 50$. Figure 6 is a plot of $\ln L(t)$ versus t using a 3σ control chart that signaled at $T = 55$. Applying the confidence set estimator in equation (3.23) with $D = 1.5$ yields a cardinality of 6. The figure shows that $t = 45$ through $t = 50$ are in the confidence set. Note the true change point, $\tau = 50$, is included in this set. A confidence set will give process engineers a window of potential change points that they can use to search for an assignable cause. Monte Carlo simulation is used to evaluate the performance of these confidence sets in Chapter 4.

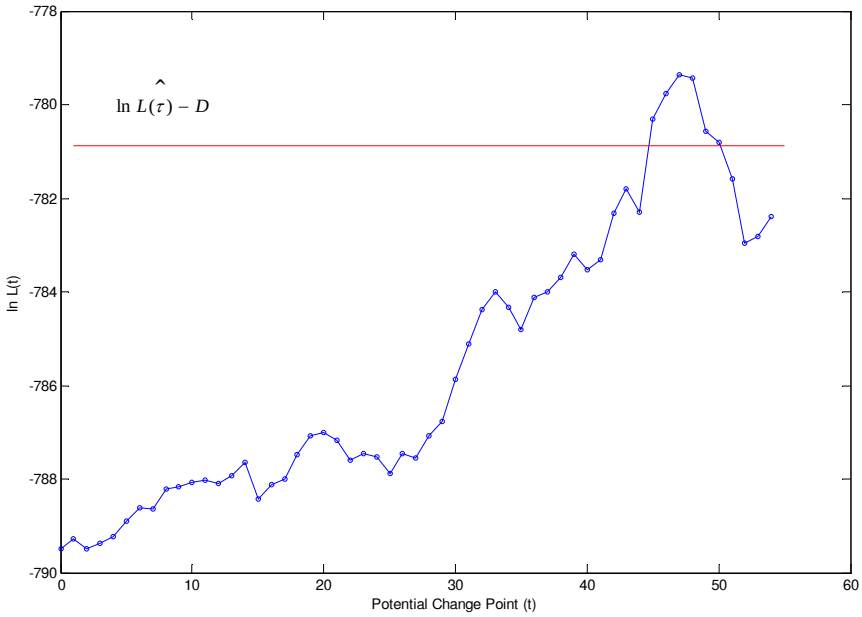

Figure 6: Log-Likelihood Plot

3.8 Conclusion

This chapter started by deriving the gamma-Poisson mixture model. This is the probability model that was the basis for modeling overdispersed count data in this research effort. In addition, two different step-change scenarios were addressed. First, the change-point model for a process that experiences a step change in mean was defined, followed by the derivation of the MLE of the process change point. Second, the change-point model for a process that experiences a step change in the overdispersion parameter was defined, followed by the derivation of the MLE for this case. In order to derive the MLE for the latter case, Newton's method was used to estimate α_a. This approach was discussed, along with the additional constraint required to keep the estimates of α_a positive. Last, the method of building a confidence set to capture the true change point was discussed. Chapter 4 uses Monte Carlo simulation to evaluate both change-point estimators proposed in this chapter. Likewise, Monte Carlo simulation is used to evaluate the performance of the confidence set estimators. Chapter 4 concludes with an application of these methods to a real-world problem.

4. Results and Analysis

In Chapter 3, a change-point estimator was derived for the time of step change in each of the parameters of the gamma-Poisson mixture model using the method of maximum likelihood estimation. This chapter discusses the Monte Carlo simulation experiments used to evaluate the root mean square error performance of the proposed estimator, $\hat{\tau}_\nu$, after genuine 3σ control chart signals. Since there is no standardized negative binomial distribution, this thesis investigates the 12 combinations of the following arbitrarily chosen in-control parameter values: ν_0 = 1, 5, and 20, and α = 1, 5, 10, and 50.

Second, this chapter discusses the Monte Carlo simulation experiments used to evaluate the root mean square error performance of the proposed estimator, $\hat{\tau}_\alpha$, after genuine CUSUM control chart signals. Since there is no standardized negative binomial distribution, this thesis investigates the 6 combinations of the following arbitrarily chosen in-control parameter values: ν = 5 and 20, and α_0 = 1, 5, and 10.

Monte Carlo simulation is also used to evaluate the performance of the confidence set estimators for each case considered. Last, this chapter applies the methodologies developed in Chapter 3 to Iraq war casualty data to demonstrate the effectiveness of these techniques on a real-world problem.

4.1 Monte Carlo Simulation When ν_a is Unknown

Monte Carlo simulation is used to evaluate the root mean square error performance of the change-point estimator following a genuine signal from a 3σ control

chart. The 3σ control chart was chosen due to its wide use and ease of implementation. Each count observed will be compared to the upper and lower control limits of the control chart. The control limits are calculated as

$$\begin{aligned} UCL &= \nu_0 + 3\sqrt{\nu_0 + \tfrac{\nu_0^2}{\alpha}} \\ LCL &= \nu_0 - 3\sqrt{\nu_0 + \tfrac{\nu_0^2}{\alpha}} \end{aligned} \tag{4.1}$$

and if LCL < 0, then set LCL = 0.

The process change point was simulated to occur at τ = 50. Independent observations were drawn from a negative binomial process where $y_i \sim \text{NB}(\alpha, \frac{\alpha}{\alpha+\nu_0})$ for $i = 1, 2, ..., 50$. Following the 50th sample collected, observations were drawn from a negative binomial process, where $y_i \sim \text{NB}(\alpha, \frac{\alpha}{\alpha+\nu_a})$ for $i = \tau + 1, ..., T$, where T is the time of control chart signal. Following the signal, $\hat{\tau}_\nu$ was calculated from equation (3.14). This procedure was repeated a total of N = 10,000 times for each value of ν_a investigated. The overdispersion parameter, α, and the in-control mean, ν_0, are both assumed known. Note that α remains unchanged after the step change occurs. The out-of-control mean, ν_a, and the change point, τ, are both assumed unknown.

4.1.1 False Alarms

As described above, each simulated observation will be compared to the upper and lower control limits of the 3σ control chart. If the control chart signals at an observation T, where $T > \tau$, then the control chart signals that the process is out-of-control. However, if the control chart signals at T, where $T < \tau$, then it is treated as a false alarm since the chart signaled before the simulated change point. In the simulation,

if a false alarm is encountered at time *T*, the control chart is restarted at time *T+1*, and the simulated change point will occur as scheduled. This is consistent with how a false alarm would be handled in practice, i.e., if a false alarm is encountered, the process engineer would declare it was a false alarm, and the process monitoring is continued at the next observation. Pignatiello and Samuel [16] handle false alarms this way.

4.2 Performance of the MLE for Step Change in Mean

The performance of the MLE, $\hat{\tau}_\nu$, for various combinations of α and ν_0 are shown in the tables throughout this section. Appendix A contains all the results from the 12 combinations of ν_0 and α studied in this thesis. Before getting to the specific results shown in the tables, column headings of the tables will be defined. For example, see Table 3 below. Recall that average run length (ARL) is defined as the expected number of observations that are observed before the control chart issues a signal. Root mean square (RMS) error is defined as the square root of mean square error (MSE). MSE, a function of both variance and bias, is defined as

$$\text{MSE} = (B(\hat{\tau}))^2 + \text{Var}(\hat{\tau}) \tag{4.2}$$

where $B(\hat{\tau}) = E(\hat{\tau}) - \tau$ is the bias of the estimator.

Table 3 also shows $\bar{\hat{\tau}}_\nu$, which is the average of the change-point estimates obtained over the *N*=10,000 independent simulation runs. The last column of the table is the standard error of $\hat{\tau}_\nu$.

Three general results can be observed from the data in this section. The most obvious is for a given ν_0 and α. It can be seen that by looking down the columns of any

individual table, say Table 3, that as the magnitude of the step change increases, ARL decreases, $\mathrm{RMS}(\hat{\tau}_\nu)$ decreases, and $\bar{\hat{\tau}}_\nu$ approaches the true value of τ. This should be intuitive in that a larger step change in mean should be estimated more accurately than a smaller step change. That is, the bias goes to zero as the magnitude of the step change increases.

Table 3: Performance of MLE for α=1, $\nu_0 = 5$, τ=50, and $N = 10{,}000$

ν_a	% Increase	ARL	$\mathrm{RMS}(\hat{\tau}_\nu)$	$\bar{\hat{\tau}}_\nu$	Std.error($\hat{\tau}_\nu$)
6	20	28.90	37.49	73.50	0.292
7	40	18.65	21.99	62.62	0.180
8	60	13.41	14.78	57.41	0.128
9	80	10.12	11.23	54.64	0.102
10	100	8.16	9.07	53.35	0.084
11	120	6.80	7.96	52.21	0.076
12	140	5.74	7.08	51.56	0.069
13	160	5.06	6.40	51.30	0.063
14	180	4.58	5.96	51.04	0.059
15	200	4.05	5.93	50.70	0.059
20	300	2.95	4.52	50.23	0.045
25	400	2.37	4.33	50.01	0.043
30	500	2.05	4.06	49.88	0.041
35	600	1.84	4.15	49.78	0.042
40	700	1.73	3.28	49.86	0.033
45	800	1.63	3.55	49.81	0.035
50	900	1.56	3.57	49.82	0.036
55	1000	1.47	3.27	49.80	0.033

The second general result that can be seen is by comparing different $\alpha's$ for a given ν_0. For example, take Table 3, Table 6, Table 7, and Table 8 and compare the summary statistics for $\nu_a = 35$. Table 4 summarizes these results as α increases for fixed values of $\nu_0 = 5$ and $\nu_a = 35$. As the variance of the observations decreases due to

an increasing α, ARL and RMS($\hat{\tau}_\nu$) both decrease, and $\bar{\hat{\tau}}_\nu$ approaches the true value of τ more quickly. Figure 7 visually shows the effect of an increasing α on RMS.

Table 4: Effect on MLE for Increasing α.

$\nu_0 = 5$ and $\nu_a = 35$. τ=50, and N=10,000

α	ARL	RMS($\hat{\tau}_\nu$)	$\bar{\hat{\tau}}_\nu$
1	1.84	4.15	49.78
5	1.08	1.93	49.81
10	1.02	1.09	49.92
50	1.00	0.08	50.00

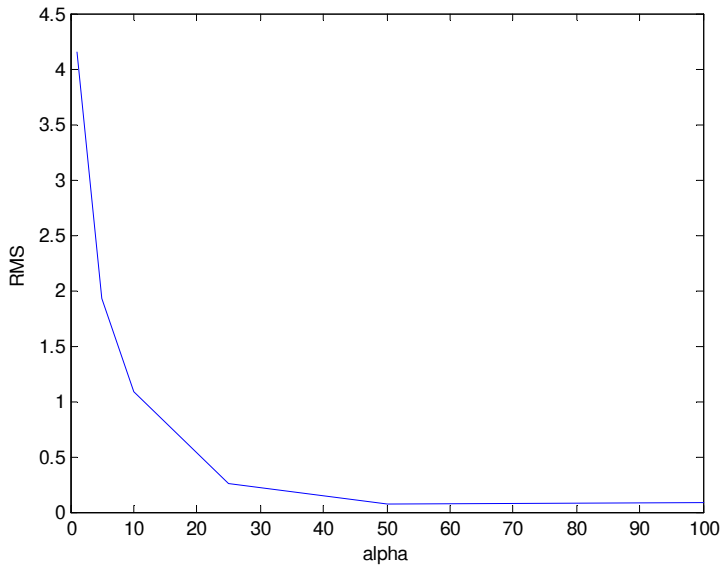

Figure 7: Effect on RMS due to an Increasing α.

$\nu_0 = 5$ and $\nu_a = 35$. τ=50, and N=10,000

A similar but more pronounced example of this second result can be seen next. When $\alpha = 50$, the estimator approaches the true change point much more quickly than when $\alpha = 1$. For example, when $\alpha = 50$ and ν_0 experiences only a 40% change,

$\bar{\hat{\tau}}_\nu$ =51.92. This is in contrast to when $\alpha = 1$ and ν_0 experiences a 40% change. In this case, $\bar{\hat{\tau}}_\nu$ =62.62. These results, extracted from Table 3 and Table 8, show the effect of alpha on the MLE and are summarized in Table 5 below. Again, the conclusion is that as the overdispersion parameter, α, increases, the variance of the observations decrease, resulting in a more accurate estimate of the process change point.

Table 5: Comparison of MLE for $\alpha = 1$ and α=50.

$\nu_0 = 5$, τ=50, and N=10,000

$\alpha = 1$	$\bar{\hat{\tau}}_\nu$	$\alpha = 50$	$\bar{\hat{\tau}}_\nu$
$\nu_a = 6$	73.50	$\nu_a = 6$	63.39
$\nu_a = 7$	62.62	$\nu_a = 7$	51.92
$\nu_a = 10$	53.35	$\nu_a = 10$	49.77

Table 6: Performance of MLE for α=5, $\nu_0 = 5$, τ=50, and $N = 10,000$

ν_a	% Increase	ARL	RMS($\hat{\tau}_\nu$)	$\bar{\hat{\tau}}_\nu$	Std.error($\hat{\tau}_\nu$)
6	20	39.01	36.96	70.98	0.304
7	40	19.45	14.16	55.62	0.130
8	60	11.39	8.58	51.83	0.084
9	80	7.36	6.22	50.67	0.062
10	100	5.28	5.59	50.03	0.056
11	120	4.12	4.75	49.92	0.048
12	140	3.31	4.40	49.71	0.044
13	160	2.78	4.40	49.57	0.044
14	180	2.41	4.00	49.61	0.040
15	200	2.15	4.05	49.50	0.040
20	300	1.49	3.62	49.55	0.036
25	400	1.25	2.57	49.70	0.026
30	500	1.14	2.46	49.71	0.024
35	600	1.08	1.93	49.81	0.019
40	700	1.06	1.62	49.84	0.016
45	800	1.03	1.39	49.89	0.014
50	900	1.02	0.96	49.93	0.010
55	1000	1.02	1.13	49.92	0.011

Table 7: Performance of MLE for α=10, $\nu_0 = 5$, τ=50, and $N = 10,000$

ν_a	% Increase	ARL	RMS($\hat{\tau}_\nu$)	$\bar{\hat{\tau}}_\nu$	Std.error($\hat{\tau}_\nu$)
6	20	51.07	35.04	68.49	0.298
7	40	22.41	11.77	53.68	0.112
8	60	11.93	7.02	50.98	0.070
9	80	7.28	5.22	50.26	0.052
10	100	4.91	4.30	49.89	0.043
11	120	3.66	4.02	49.69	0.040
12	140	2.86	3.80	49.64	0.038
13	160	2.36	3.78	49.53	0.038
14	180	2.07	3.35	49.60	0.033
15	200	1.79	3.37	49.54	0.033
20	300	1.26	2.71	49.65	0.027
25	400	1.11	2.08	49.78	0.021
30	500	1.04	1.30	49.89	0.013
35	600	1.02	1.09	49.92	0.011
40	700	1.01	0.79	49.95	0.008
45	800	1.00	0.51	49.97	0.005
50	900	1.00	0.47	49.98	0.005
55	1000	1.00	0.70	49.98	0.007

Table 8: Performance of MLE for α=50, $\nu_0 = 5$, τ=50, and $N = 10,000$

ν_a	% Increase	ARL	RMS($\hat{\tau}_\nu$)	$\bar{\hat{\tau}}_\nu$	Std.error($\hat{\tau}_\nu$)
6	20	75.77	30.75	63.39	0.277
7	40	27.62	9.54	51.92	0.094
8	60	12.84	5.63	50.42	0.056
9	80	7.12	4.27	49.92	0.043
10	100	4.45	3.54	49.77	0.035
11	120	3.14	3.36	49.65	0.033
12	140	2.38	3.26	49.60	0.032
13	160	1.91	3.11	49.59	0.031
14	180	1.63	3.01	49.59	0.030
15	200	1.41	2.67	49.64	0.027
20	300	1.08	1.61	49.82	0.016
25	400	1.01	0.95	49.93	0.010
30	500	1.00	0.33	49.98	0.003
35	600	1.00	0.08	50.00	0.001
40	700	1.00	0.05	50.00	0.001
45	800	1.00	0.02	50.00	0.000
50	900	1.00	0.00	50.00	0.000
55	1000	1.00	0.01	50.00	0.000

The simulation results also show that the MLE is more accurate for larger in-control values of ν_0 for a given α and ν_a. For example, look at Table 8 along with Table 10 and Table 11. It can be seen for a fixed α, and a larger ν_0, the change point is estimated much more accurately for a given percent change in mean. For example, a 40% increase in $\nu_0 = 1$ is $\nu_a = 1.4$, which gives an MLE of $\bar{\hat{\tau}}_\nu = 115.84$. Contrast this to a 40% increase in $\nu_0 = 20$ which is $\nu_a = 28$, which gives an MLE of $\bar{\hat{\tau}}_\nu = 52.56$. Note that these results hold due to the fact that the out-of-control mean is specified as a percentage increase of the in-control mean. The conclusion here is that a larger shift in mean will result in a more accurate estimate of the true change point. In some instances, in practice, smaller shifts (e.g. 10% or 20%) may not be as important to detect as large shifts. Table 9 compares the MLE for different in-control values of ν_0 for the given percent changes shown.

Table 9: Comparison of MLE for $\alpha = 50$ for $\nu_0 = 1,\ 5$, and 10.

τ=50, and N=10,000

% Increase	$\nu_0 = 1$	$\nu_0 = 5$	$\nu_0 = 20$
20	115.84	63.39	52.56
40	67.90	51.92	49.94
100	51.84	49.77	49.60
400	49.60	49.93	50.00

Table 10: Performance of MLE for α=50, $\nu_0 = 1$, τ=50, and $N = 10,000$

ν_a	% Increase	ARL	RMS($\hat{\tau}_\nu$)	$\bar{\hat{\tau}}_\nu$	Std.error($\hat{\tau}_\nu$)
1.2	20	117.43	103.96	115.84	0.805
1.4	40	64.79	35.47	67.90	0.306
1.6	60	39.11	18.26	57.18	0.168
1.8	80	25.67	11.56	53.28	0.111
2	100	17.80	8.47	51.84	0.083
2.2	120	13.02	6.53	51.05	0.064
2.4	140	9.96	5.52	50.57	0.055
2.6	160	7.81	5.04	50.29	0.050
2.8	180	6.35	4.68	50.06	0.047
3	200	5.26	4.71	49.87	0.047
4	300	2.68	3.53	49.65	0.035
5	400	1.81	3.14	49.60	0.031
6	500	1.44	2.93	49.65	0.029
7	600	1.23	2.61	49.67	0.026
8	700	1.13	1.87	49.78	0.019
9	800	1.07	1.60	49.83	0.016
10	900	1.04	1.53	49.86	0.015
11	1000	1.02	1.29	49.89	0.013

Table 11: Performance of MLE for α=50, $\nu_0 = 20$, τ=50, and $N = 10,000$

ν_a	% Increase	ARL	RMS($\hat{\tau}_\nu$)	$\bar{\hat{\tau}}_\nu$	Std.error($\hat{\tau}_\nu$)
24	20	28.56	10.87	52.56	0.106
28	40	7.76	4.54	49.94	0.045
32	60	3.33	3.50	49.61	0.035
36	80	1.96	3.13	49.56	0.031
40	100	1.45	2.97	49.60	0.029
44	120	1.21	2.57	49.66	0.026
48	140	1.10	2.37	49.73	0.024
52	160	1.04	1.63	49.84	0.016
56	180	1.02	1.10	49.91	0.011
60	200	1.01	0.76	49.94	0.008
80	300	1.00	0.12	49.99	0.001
100	400	1.00	0.05	50.00	0.001
120	500	1.00	0.00	50.00	0.000
140	600	1.00	0.00	50.00	0.000
160	700	1.00	0.00	50.00	0.000
180	800	1.00	0.00	50.00	0.000
200	900	1.00	0.00	50.00	0.000
220	1000	1.00	0.00	50.00	0.000

General results indicate that as the magnitude of the step change increases, or as the variance of the observations is lowered, the true change point can be estimated more accurately in a process that experiences a step change in mean. The above results show that the proposed MLE performs well. There are some exceptions for small step changes with highly variable data, but overall, the true change point is still estimated very accurately.

The importance of having an accurate method for estimating the true change in a process should now be more obvious. Look at Table 8 where $\nu_0 = 5,\ \nu_a = 7,\ \alpha$=50, and ARL=27.62. The estimate of the change point would be badly biased if an engineer estimated the time of the process change at the time the control chart signaled. It is possible that the process engineer could incorrectly diagnose the root cause of the change or perhaps not even discover the root cause at all. Instead of searching for causes around the correct time of change, a process engineer might examine log books and records associated with a time frame that was almost 28 observations after the process actually changed.

4.3 Cardinality and Coverage Performances of Confidence Set Estimators for Step Change in Mean

Set cardinality and coverage measures are used to evaluate the confidence set estimators for the nine values of *D*, where *D*=1, 1.25, 1.50,…, 3. Recall from equation (3.23) that *D* is a specified constant used to give a desired confidence and cardinality for each percent increase in ν relative to ν_0. The change point was evaluated at 20% increases in the mean (relative to ν_0) up to a 200% increase. Following this, the change

point was evaluated at 100% increases in the mean up to a 1000% increase. Recall that the process change point was simulated to occur at τ = 50. After the 3σ control chart signaled a genuine step change, the confidence set estimator was applied. The cardinality of the confidence set was recorded as well as whether the true change point was included in that set. This procedure was repeated for a total of $N = 10,000$ independent simulation runs for each percent increase in ν relative to ν_0. The average cardinality was computed over the 10,000 runs. The proportion of the 10,000 runs that included the true change point in the estimated set, i.e., coverage, was also computed.

Surface plots were created for the 12 combinations of α and ν_0. These plots will aid the user in selecting a value of *D* that meets his or her desired level of confidence. This is the same approach used by Perry and Pignatiello [12]. While all 12 plots are shown in Appendix B, four of the surface plots are shown below.

The figures show that more coverage can be obtained for a given percent increase in ν relative to ν_0 by selecting a larger reference value *D*. The tradeoff is that the confidence set will also have a larger cardinality. For example in Figure 8, a 40% change in ν for *D* = 1.5 gives a coverage of 0.39 and cardinality of 16, whereas *D* = 2 gives a coverage of 0.64 and cardinality of 35. For the case where *D* = 2, the interpretation is that the user can be 64% confident that the true change point will be contained in the set of 35 points.

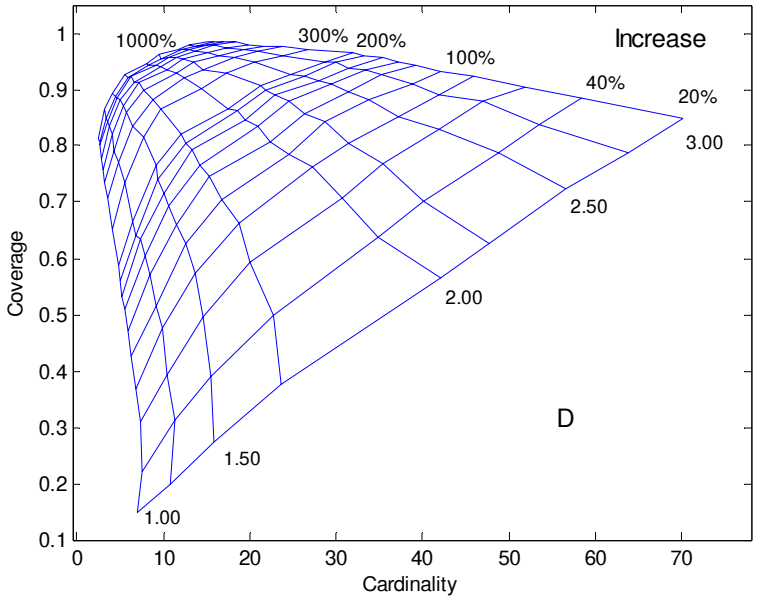

Figure 8: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Set Cardinality, Coverage, Percent Increase from ν_0, and Reference Value D. $\alpha = 1,\ \nu_0 = 1,\ \tau = 50$, and $N = 10,000$.

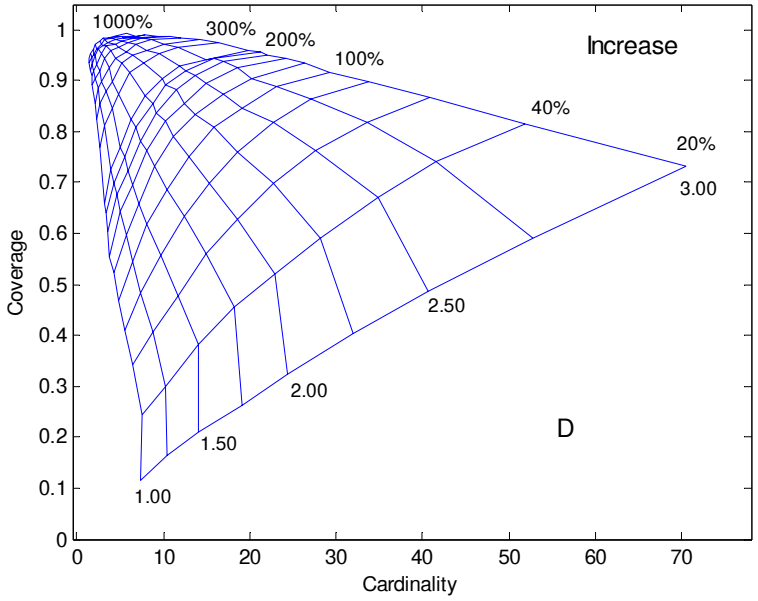

Figure 9: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Set Cardinality, Coverage, Percent Increase from ν_0, and Reference Value D. $\alpha = 5,\ \nu_0 = 1,\ \tau = 50$, and $N = 10,000$.

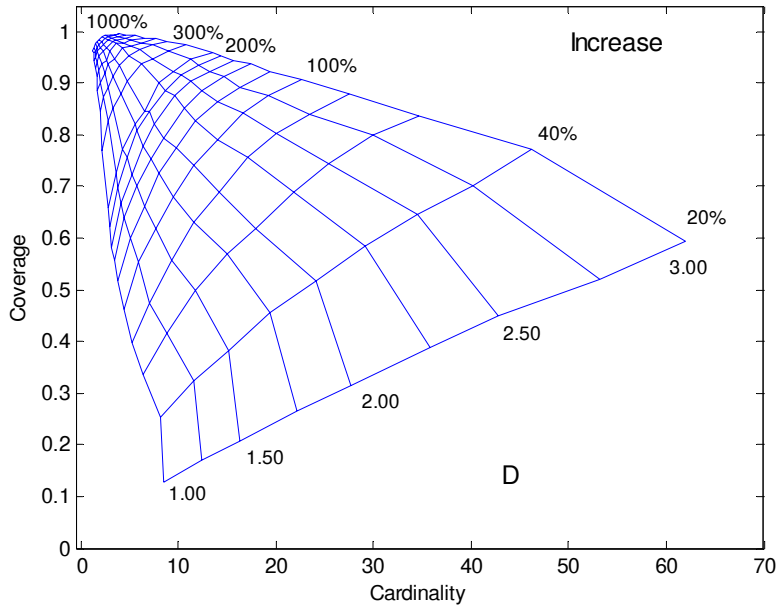

Figure 10: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Cardinality, Coverage, Percent Increase from ν_0, and Reference Value D. $\alpha = 10,\ \nu_0 = 1,\ \tau = 50$, and $N = 10,000$.

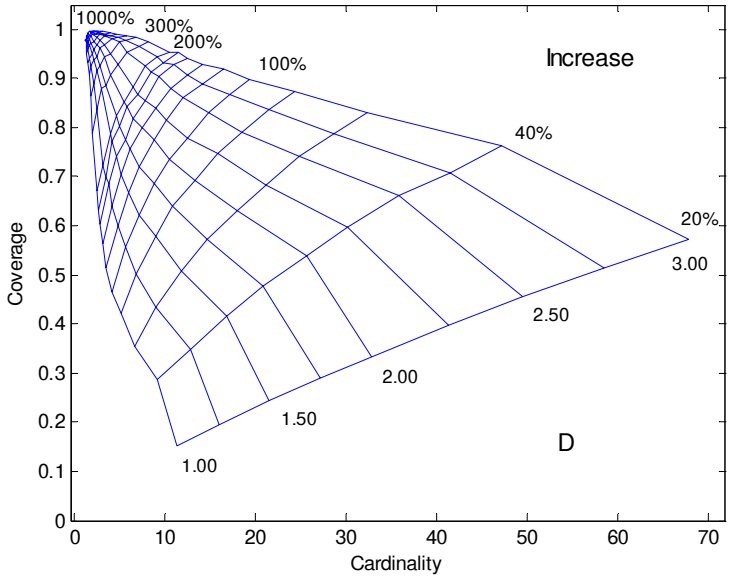

Figure 11: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Cardinality, Coverage, Percent Increase from ν_0, and Reference Value D. $\alpha = 50,\ \nu_0 = 1,\ \tau = 50$, and $N = 10,000$.

4.3.1 Choice of τ for Performance Evaluation of the Confidence Set Estimator

For this research, the process change point was simulated to occur at τ = 50. Results in Table 12 show that the choice of τ in this simulation study has little impact on the coverage probabilities. Cardinality, however, appears to increase slightly for increases in τ, as shown in Table 13. While different values of τ could have been used for this simulation study, overall the choice of $\tau = 50$ seems reasonable. These finding are consistent with results published in Pignatiello and Samuel [16].

Table 12: Effect of the Change in τ on Coverage Probability.

$\alpha = 5,\ \nu_0 = 20,$ and $N = 10,000$

% Increase	$\tau = 1$	$\tau = 50$	$\tau = 100$
20	0.33	0.28	0.27
80	0.69	0.64	0.63
160	0.83	0.81	0.81
300	0.93	0.91	0.91

Table 13: Effect of the Change in τ on Average Size of Confidence Sets.

$\alpha = 5,\ \nu_0 = 20,$ and $N = 10,000$

% Increase	$\tau = 1$	$\tau = 50$	$\tau = 100$
20	5.42	7.73	8.26
80	2.28	3.17	3.47
160	1.59	2.19	2.42
300	1.29	1.67	1.72

4.4 Monte Carlo Simulation When α_a is Unknown

Monte Carlo simulation is used to evaluate the root mean square error performance of the change-point estimator following a signal from a CUSUM control chart. A CUSUM was chosen for this case to simply add variety to this research effort. The derivation of this CUSUM is in section 4.5.

The process change point was simulated to occur at τ = 50. Independent observations were drawn from a negative binomial process where $y_i \sim \text{NB}(\alpha_0, \frac{\alpha_0}{\alpha_0+\nu})$, for $i = 1, 2, ..., 50$. Following the 50th sample collected, observations were drawn from a negative binomial process, where $y_i \sim \text{NB}(\alpha_a, \frac{\alpha_a}{\alpha_a+\nu})$, for $i = \tau + 1$, ..., T, and T denotes the time of control chart signal. Following the detection signal, $\hat{\tau}_\alpha$ is calculated from equation (3.20). This procedure was repeated a total of N = 10,000 times for each value of α_a investigated. Since there is no standardized negative binomial distribution, this thesis investigated the 6 combinations of the following arbitrarily chosen in-control parameter values: ν = 5 and 20, and α_0 = 1, 5, and 10. The in-control overdispersion parameter, α_0, and the process mean, ν, are both assumed known, where ν remains constant. The out-of-control overdispersion parameter, α_a, and the change point, τ, are unknown. False alarms were handled exactly as described previously.

4.5 CUSUM Control Chart for Detecting a Change in Overdispersion Parameter

For this research, the CUSUM test statistic for detecting a decrease in the overdispersion parameter is given by $C_i^- = \max\{C_i^-, 0\}$, where C_i^- is the cumulative sum at time *i*. Evidence of a decrease in the overdispersion parameter is indicated by a signal occurring if $C_i^- > h^-$. A one-sided tabular CUSUM was designed to quickly detect a 25% decrease in α relative to α_0 with the in-control average run length (ARL_0) calibrated to approximately 370.

The CUSUM control chart used to detect a decrease from α_0 was derived from the sequential probability ratio test (SPRT). See Hawkins and Olwell [6] for details of using the SPRT to derive a CUSUM statistic.

In the SPRT, the null hypothesis $H_0: \alpha_i = \alpha_0,\ i = 1, 2, ..., T$ is tested against $H_a: \alpha_i = \alpha_a^*,\ i = 1, 2, ..., T$. In SPRT , the likelihood ratio is given by

$$Z_i^- = \frac{\prod_{i=1}^{T} \frac{\Gamma(y_i + \alpha_a^*)}{\Gamma(y_i + 1)\Gamma(\alpha_a^*)} \left(\frac{\nu}{\alpha_a^* + \nu}\right)^{y_i} \left(\frac{\alpha_a^*}{\alpha_a^* + \nu}\right)^{\alpha_a^*}}{\prod_{i=1}^{T} \frac{\Gamma(y_i + \alpha_0)}{\Gamma(y_i + 1)\Gamma(\alpha_0)} \left(\frac{\nu}{\alpha_0 + \nu}\right)^{y_i} \left(\frac{\alpha_0}{\alpha_0 + \nu}\right)^{\alpha_0}} \tag{4.3}$$

where the value α_0 is the in-control value of α, the value α_a^* is the pre-specified out-of-control value of α that one wishes to detect, and ν is the process mean, assumed constant.

To derive the CUSUM statistic, take the natural log of equation (4.3) and simplify to obtain

$$C_i^- = \sum_{i=1}^{T} \left[\ln\left(\frac{\Gamma(y_i + \alpha_a^*)\Gamma(\alpha_0)}{\Gamma(y_i + \alpha_0)\Gamma(\alpha_a^*)}\right) + \ln\left(\frac{\alpha_0 + \nu}{\alpha_a^* + \nu}\right)^{y_i} + \alpha_a^* \ln\left(\frac{\alpha_a^*}{\alpha_a^* + \nu}\right) - \alpha_0 \ln\left(\frac{\alpha_0}{\alpha_0 + \nu}\right) \right]. \tag{4.4}$$

Rewriting this expression gives

$$C_i^- = \ln\left(\frac{\Gamma(y_T + \alpha_a^*)\Gamma(\alpha_0)}{\Gamma(y_T + \alpha_0)\Gamma(\alpha_a^*)}\right) + \ln\left(\frac{\alpha_0 + \nu}{\alpha_a^* + \nu}\right)^{y_T} + \alpha_a^* \ln\left(\frac{\alpha_a^*}{\alpha_a^* + \nu}\right) - \alpha_0 \ln\left(\frac{\alpha_0}{\alpha_0 + \nu}\right) + C_{i-1}^- \tag{4.5}$$

where the value of C_0^- is taken to be zero. Equation (4.5) is set to zero whenever $C_i^- < 0$. The SPRT operates by comparing C_i^- to the decision interval h^- at each new observation. If $C_i^- > h^-$, then the test concludes in favor of H_a.

Note that the familiar tabular form of the CUSUM statistic is not ready available since the observed count is contained in the gamma function. However, since this is a SPRT, it still remains a CUSUM.

4.6 Performance of the MLE for Step Change in the Overdispersion Parameter

Three general results can be observed from the data in this section. The most obvious is for a given α_0 and ν. It can be seen that by looking down the columns of any of these tables, say Table 14, that as the magnitude of step change increases, ARL and RMS($\hat{\tau}_\alpha$) both decrease, and $\bar{\hat{\tau}}_\alpha$ approaches the true value of τ. The conclusion is that a larger magnitude of step change in α will result in a more accurate estimate of the change point, i.e., the bias goes to zero as the magnitude of the step change increases.

Table 14: Performance of MLE for $\nu = 5,\ \alpha_0 = 10,\ \tau = 50,$ and N=10,000

α_a	% Decrease	ARL	RMS($\hat{\tau}_\alpha$)	$\bar{\hat{\tau}}_\alpha$	Std.error($\bar{\hat{\tau}}_\alpha$)
9	10	246.97	327.69	276.17	2.371
8	20	175.21	226.82	205.20	1.654
7	30	124.48	153.54	155.15	1.119
6	40	85.02	98.24	115.80	0.730
5	50	59.27	63.04	90.23	0.485
4	60	40.55	39.48	72.81	0.322
3	70	26.70	23.51	60.72	0.209
2	80	16.92	13.65	53.77	0.131
1	90	9.26	7.75	50.18	0.078

The second general result that can be seen is by comparing different $\nu's$ for a given α_0 and given percent decreases. For example, take Table 14 and Table 19 and compare the performance of $\hat{\tau}_\alpha$ for $\nu = 5$ and $\nu = 20$. Table 15 shows that when $\nu = 20$, the estimator approaches the true change point more quickly than when $\nu = 5$.

The conclusion is that a larger ν will experience a larger magnitude of change compared to a smaller ν for a given α_0 and given percent decrease. Again, the bias goes to zero as the magnitude of the step change increases. Note that the results in this section hold due to the fact that the out-of-control overdispersion parameter is specified as a percentage decrease of the in-control parameter.

Table 15: Comparison of MLE for $\nu = 5$ and ν=20 for Various Decreases in α Relative to α_0. $\alpha_0 = 10$, τ=50, and $N = 10,000$

ν	20% Decrease	50% Decrease	80% Decrease
5	205.20	90.23	53.77
20	144.19	64.21	50.93

The third general result from this section is that the MLE is more accurate for smaller in-control values of α_0 for a given mean and percent decrease in α relative to α_0. For example, look at Table 17, Table 18, and Table 19 below. It can be seen for a fixed ν and given percent decreases in α (relative to α_0), the change point is estimated more accurately at a lower α_0. For example, consider a 50% decrease from $\alpha = 10$ to $\alpha = 5$, which gives an MLE of $\bar{\hat{\tau}}_\alpha = 64.21$. Contrast this to a 50% decrease from $\alpha = 1$, to $\alpha = 0.5$. This case gives an MLE of $\bar{\hat{\tau}}_\alpha = 56.31$. The conclusion is that a smaller α_0 causes higher variance in the observations, *y*, for a given mean and percent decrease. It appears that this increased magnitude in variance allows $\hat{\tau}_\alpha$ to be estimated more accurately. Note that the variance of *y*, as shown in equation (3.9), can be made larger in two ways. The first way to increase the variance is by increasing ν for a given α. The

second way the variance can be made larger is to decrease α for a given ν, which is what was just discussed. These results are summarized in Table 16.

Table 16: Comparison of MLE for $\nu = 20$ for $\alpha_0 = 1,\ 5,$ and 10.

τ=50, and $N = 10,000$

% Decrease	$\alpha_0 = 1$	$\alpha_0 = 5$	$\alpha_0 = 10$
20%	109.05	126.01	144.19
50%	56.31	60.60	64.21
80%	49.51	50.51	50.93

Table 17: Performance of MLE for $\nu = 20,\ \alpha_0 = 1,\ \tau = 50,$ and N=10,000

α_a	% Decrease	ARL	RMS($\hat{\tau}_\alpha$)	$\bar{\hat{\tau}}_\alpha$	Std.error($\bar{\hat{\tau}}_\alpha$)
0.9	10	151.83	198.58	184.50	1.461
0.8	20	76.51	89.16	109.05	0.668
0.7	30	43.77	46.07	77.89	0.367
0.6	40	27.73	25.38	63.51	0.215
0.5	50	18.87	15.70	56.31	0.144
0.4	60	13.30	10.27	52.61	0.099
0.3	70	9.46	7.22	50.53	0.072
0.2	80	6.77	5.68	49.51	0.057
0.1	90	4.76	4.57	48.98	0.045

Table 18: Performance of MLE for $\nu = 20,\ \alpha_0 = 5,\ \tau = 50,$ and N=10,000

α_a	% Decrease	ARL	RMS($\hat{\tau}_\alpha$)	$\bar{\hat{\tau}}_\alpha$	Std.error($\bar{\hat{\tau}}_\alpha$)
4.5	10	177.80	235.03	209.76	1.724
4	20	93.53	114.08	126.01	0.851
3.5	30	55.23	61.16	88.53	0.475
3	40	35.67	36.15	70.57	0.297
2.5	50	23.46	21.45	60.60	0.186
2	60	15.53	13.51	54.73	0.127
1.5	70	10.52	9.17	52.00	0.090
1	80	6.77	6.23	50.51	0.062
0.5	90	3.99	4.11	49.86	0.041

Table 19: Performance of MLE for $\nu = 20,\ \alpha_0 = 10,\ \tau = 50,$ and N=10,000

α_a	% Decrease	ARL	RMS($\hat{\tau}_\alpha$)	$\bar{\hat{\tau}}_\alpha$	Std.error($\bar{\hat{\tau}}_\alpha$)
9	10	192.08	252.41	223.78	1.831
8	20	112.28	140.27	144.19	1.039
7	30	66.85	76.22	99.57	0.579
6	40	43.35	45.26	77.37	0.360
5	50	28.19	26.67	64.21	0.226
4	60	19.11	16.79	57.04	0.152
3	70	12.62	10.87	52.88	0.105
2	80	8.02	7.47	50.93	0.074
1	90	4.44	4.76	50.01	0.048

General results indicate that as the magnitude of the step change increases, the true change point can be estimated more accurately. On average, the results show that the proposed MLE performs well. There are some exceptions as were shown, but overall, the true change point is still estimated very accurately. See Appendix C for all cases investigated.

4.7 Cardinality and Coverage Performances of Confidence Set Estimators for Step Change in Overdispersion Parameter

Set cardinality and coverage measures are used to evaluate the confidence set estimators for the nine values of *D*, where *D*=1, 1.25, 1.50,…, 3. Recall from equation (3.23) that *D* is a specified constant used to provide a desired confidence level and cardinality for each percent decrease in α relative to α_0. The change point was evaluated at 10% decreases from α_0 up to a 90% decrease. As before, the process change point was simulated to occur at τ = 50. Following a genuine CUSUM control chart signal, the confidence set estimator was applied. The cardinality of the confidence set was recorded as well as whether the true change point was included in that set. This procedure was repeated for a total of $N = 10,000$ independent simulation runs for each

percent increase in α (relative to α_0). The average cardinality was computed over the 10,000 runs. The proportion of the 10,000 runs that included the true change point in the estimated set, i.e., coverage, was also computed.

Surface plots were created for the six combinations of ν and α_0 looked at in this thesis. The plots of various *D* values will aid the user in selecting a *D* value that meets his or her desired level of confidence. While all six plots are shown in Appendix D, two of the surface plots are shown below.

As before, the figures show that more coverage can be obtained for a given percent decrease in α relative to α_0 by selecting a larger reference value *D*. The tradeoff is that the confidence set will also have a larger cardinality. For example in Figure 12, a 50% decrease from α_0 for $D = 1.5$ gives a coverage of 0.48 and cardinality of 27, whereas $D = 2$ gives a coverage of 0.64, but a cardinality of 42. For the case where $D = 2$, the interpretation is that the user can be 64% confident that the true change point will be contained in the set of 42 points.

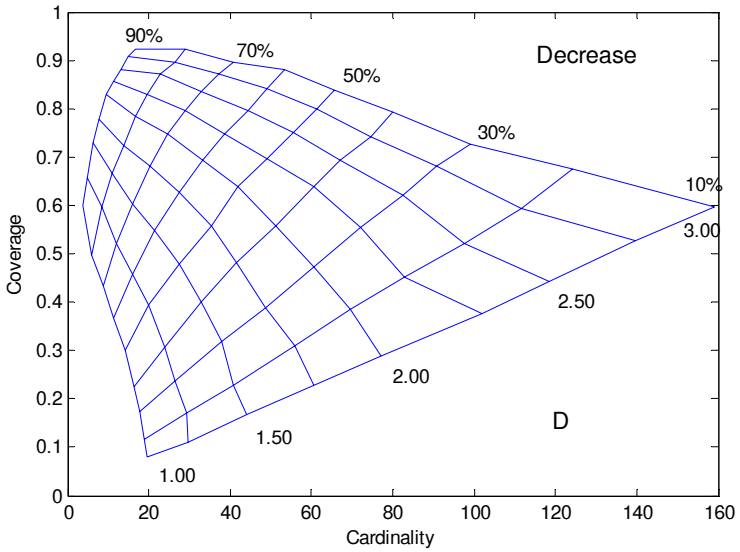

Figure 12: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Cardinality, Coverage, Percent Decrease from α_0, and Reference Value D. ν=5, α_0=5, τ=50, and $N = 10,000$.

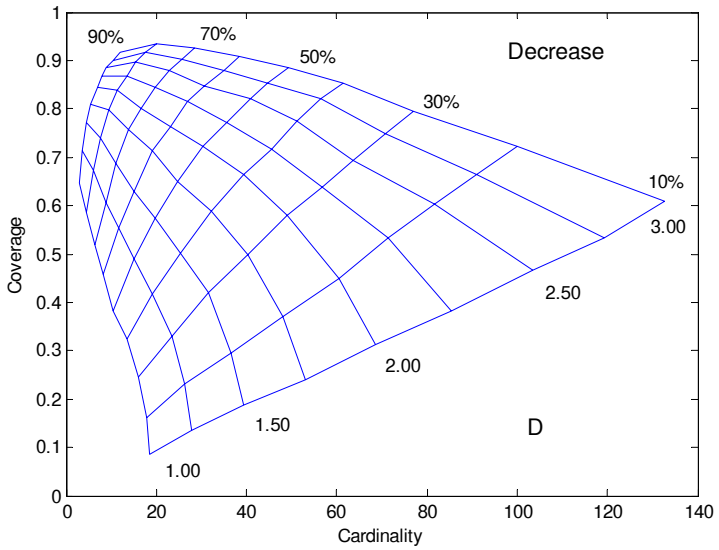

Figure 13: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Cardinality, Coverage, Percent Decrease from α_0, and Reference Value D. ν=20, α_0=5, τ=50, and $N = 10,000$.

4.7.1 Choice of τ for Performance Evaluation of the Confidence Set Estimator

For this research, the process change point was simulated to occur at τ = 50. Results in Table 20 show that the choice of τ in this simulation study has little impact on the coverage probabilities. Cardinality, however, appears to increase slightly for increases in τ, as shown in Table 21. While different values of τ could have been used for this simulation study, overall the choice of $\tau = 50$ seems reasonable. These finding are consistent with the results published in Pignatiello and Samuel [16].

Table 20: Effect of the Change in τ on Coverage Probability.

$\alpha = 5,\ \nu_0 = 20,$ and $N = 10,000$

% Decrease	$\tau = 1$	$\tau = 50$	$\tau = 100$
10	0.10	0.09	0.09
50	0.45	0.38	0.39
90	0.71	0.65	0.65

Table 21: Effect of the Change in τ on Average Size of Confidence Sets.

$\alpha = 5,\ \nu_0 = 20,$ and $N = 10,000$

% Decrease	$\tau = 1$	$\tau = 50$	$\tau = 100$
10	17.39	18.62	19.17
50	7.31	10.57	11.47
90	1.91	2.83	2.99

4.8 Change-Point Analysis Applied to Iraq War Coalition Casualty Data

The data and results of this thesis have been obtained from simulation experiments. Hence, it begs the question, "How are these techniques applied to a real-life data set?" To answer this question, a data set containing Iraq war coalition casualties was obtained from the Iraq Coalition Casualty Count website at http://icasualties.org/oif/. A realistic scenario using that data is as follows.

The US commander in Iraq has tasked the strategic assessments team to monitor coalition casualties and to report if there are any noticeable changes from the average daily coalition casualties. The assessment team decided they would use a 3σ control chart to monitor for changes, and that they would also like to apply change-point estimation techniques after the control chart signals a genuine change. Not knowing exactly what kind of change might occur; they decided to employ a step change model.

Because daily casualties are data counts, and it is suspected that the data would be overdispersed due to the variability associated with war, the gamma-Poisson mixture model was used as the underlying probability distribution of the counts. The assessment team was to start monitoring on 07/19/2003. To get a baseline, the equivalent of the phase I SPC calibration process, the assessment team took prior data from a 100-day period from 04/10/2003 – 07/18/2003. Recall that the gamma-Poisson mixture model, as parameterized in this thesis, is the negative binomial probability distribution with parameters $r = \alpha$ and $p = \frac{\alpha}{\alpha+\nu}$. After some careful analysis, the team estimated the in-control parameters as $\hat{\alpha} = 2.21$ and $\hat{\nu}_0 = 1.25$. They established upper and lower control limits as UCL = 5.4 and LCL = 0 respectively. Any daily coalition casualty count of 6 or

greater would cause the control chart to signal. They were now ready to use the control chart to monitor daily coalition casualties.

Figure 14 shows the results of the strategic assessment team's change-point analysis. The team started monitoring daily coalition casualties starting on 07/19/2003 (day 1). The control chart detected a change on 10/26/2003, or day 100 from when the daily monitoring was started. After the control chart signaled, they applied the MLE from equation (3.14) and estimated the change point as day 99, as shown graphically in Figure 15. Recall that the change point is the last point from the in-control process. In this case, the control chart signal and the estimated change point are in exact agreement.

The log-likelihood plot of Figure 15 is important in the fact that it can be used to search for assignable causes on other likely days. Even though the MLE of the change point was day 99, it can be seen from Figure 15 that potentially any of the change points from, say, days 83-99 can be investigated for assignable causes. As an example of the usefulness of the confidence set estimator, the surface plot of Figure 8 is used as a guide to narrow down the pool of potential change points even further. By setting D = 1.75 from equation (3.23), the assessment team can narrow their search window for assignable causes to about 12 days and still achieve almost a 90% confidence.

After the control chart signal, there are two conclusions that can be made. Either this signal is a false alarm and nothing is out of the ordinary, or that this is indeed a genuine signal and it warrants further investigation. The team decided to investigate further and they discovered that the Muslim holy month of Ramadan begins on 10/27/2003, or day 101 from when daily monitoring began. Their conclusion was that

this is not a false alarm and they briefed their results to the US commander in charge so that appropriate actions could be taken.

Figure 16 is insightful in that it shows what happened after the change was detected in Figure 14. In Figure 14, even though a change was detected, the daily casualty does not seem to be too far out of the ordinary. But as Figure 16 shows, the control chart indeed detected a significant change that is clearly shown by the dramatic increase in casualties during the month of Ramadan. Had the proposed method actually been used, tighter security measures could have been implemented following the control chart signal, and perhaps coalition lives could have been saved.

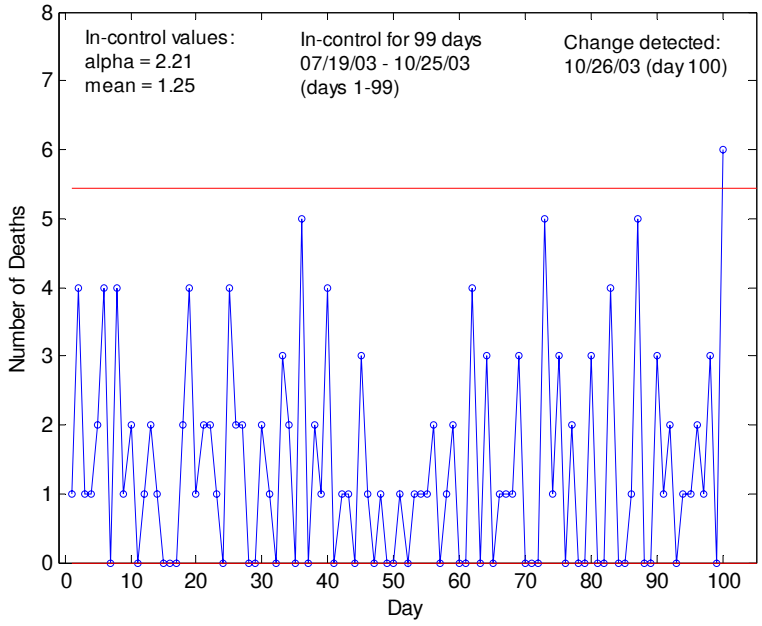

Figure 14: Control Chart, Iraq Coalition Casualties 07/19/03 – 10/26/03

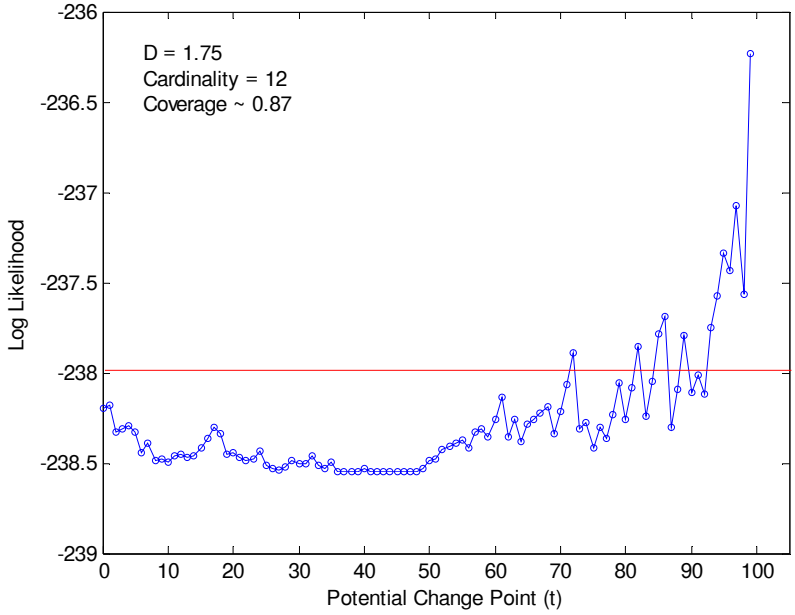

Figure 15: Log-Likelihood Plot of Iraq War Casualty Data

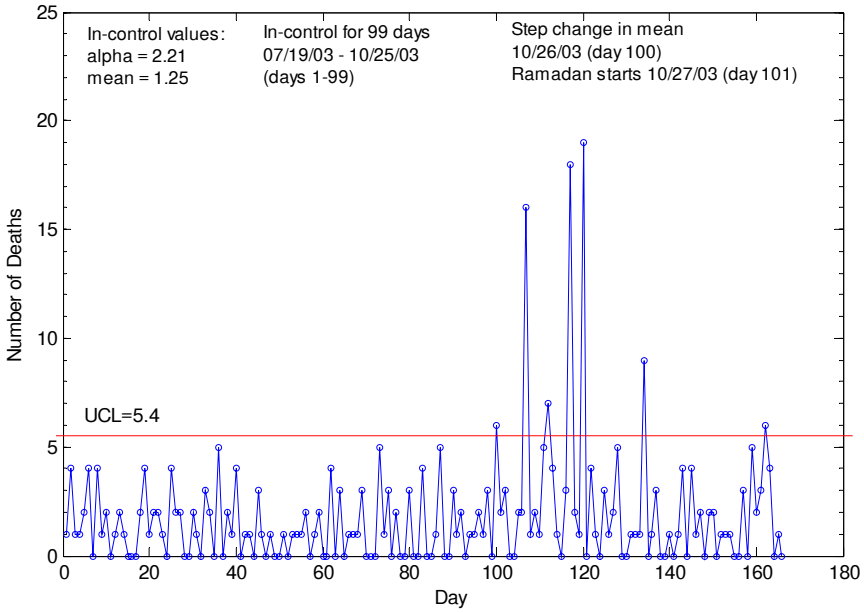

Figure 16: Coalition Casualties 07/19/03 – 12/31/03

This example was based off of real-world data. A 3σ control chart was used to monitor a process where the data was overdispersed count data. This example shows the power of using change-point methods in conjunction with a control chart to monitor a process. The control chart will allow a change to be detected, while change-point estimation techniques can narrow down the search for an assignable cause. In this example, however, the estimated change point and control chart signal coincided largely due to a spike in fatality counts. While most of the literature deals with examples from a manufacturing context, this example shows that these methods can be applied to non-traditional, non-manufacturing processes, including military processes.

4.9 Conclusion

This chapter employed Monte Carlo simulation to evaluate the performance of each of the MLEs derived in Chapter 3. The first case considered was when a count process experienced a step change in the mean. Monte Carlo simulation was used to evaluate the root mean square error performance of the change-point estimator following a signal from a 3σ control chart. General results indicate that as the magnitude of the step change increases, or as the variance of the observations is lowered, the true change point can be estimated more easily in a process that experiences a step change in mean.

The second case considered was when a count process experienced a step change in the overdispersion parameter. Monte Carlo simulation was used to evaluate the root mean square error performance of this change-point estimator following a signal from a CUSUM control chart. General results indicate that as the magnitude of the step change increases, the true change point can be estimated more accurately.

The performance of the confidence set estimator for each case was also investigated. The surface plots generated will aid a user in selecting a value for D that meets his or her desired level of confidence. Last, the methodologies developed in Chapter 3 were successfully applied to Iraq war coalition casualty data to demonstrate the effectiveness of these techniques on a real-world problem. Next, Chapter 5 will summarize this research effort and make recommendations for future research.

5. Conclusions and Future Research

5.1 Summary and Conclusions

If a process engineer could determine the actual time of a process change, their time and resources could be better used to investigate the assignable causes that contributed to the change. A quick identification of the assignable causes would lead to an improvement of the process sooner. Estimating the true change point is also important due to the costs associated with misdiagnosing a control chart signal. These costs include the time and resources used to track down any assignable causes, as well as any costs associated with unnecessary process adjustments. Therefore, the primary focus of this research was to derive and evaluate maximum likelihood estimators for the time of a step change in each of the parameters of the gamma-Poisson mixture model, following a genuine control chart signal.

To motivate a methodology for this research, a review of the relevant literature was presented. First was a review of a step change, change-point model. Then, the relevant literature on change-point estimation was discussed. That is, processes that experience step changes, linear trends, and monotonic changes have been studied for normal, Poisson, and/or binomial processes. Many cases of change-point estimation are considered in the literature; however, the case of estimating a change point, where the observations are overdispersed count data, has not received much, if any, attention in the quality engineering literature.

The methodology focused on the mathematical models employed in this research. The gamma-Poisson mixture model was derived and employed as the probability model

for modeling overdispersed count data in this thesis. Then, two different step-change scenarios were addressed. First, the change-point model for a process that experiences a step change in mean was defined, followed by the derivation of the MLE of the process change point. Second, the change-point model for a process that experiences a step change in the overdispersion parameter of the underlying probability distribution was defined. Then, the corresponding MLE of the process change point was derived for this case. Last, the method of building a confidence set to capture the true process change point was discussed. Monte Carlo simulation was used to evaluate each of these proposed estimators. Likewise, Monte Carlo simulation was used to evaluate the performance the proposed confidence set estimators.

The first case considered a step change in the mean of a process. The conclusions reached were as the magnitude of the step change increased, or as the variance of the observations was lowered, the true change point was estimated more accurately. Then set cardinality and coverage measures were used to evaluate the confidence set estimators. The given surface plots show more coverage can be obtained for a given percent increase in ν from ν_0 by selecting a larger reference value *D*, but the tradeoff is that the confidence set will also have a larger cardinality.

Second, Monte Carlo simulation was used to evaluate the performance of the proposed estimator after a CUSUM control chart detected a step change in the overdispersion parameter. Results indicate that as the magnitude of the step change increased, the true change point could be detected more accurately. The set cardinality and coverage measures were also used to evaluate the confidence set estimators in this case, with the similar findings as the step change in mean case.

To demonstrate the effectiveness of these techniques on a real-world problem, the methodology of Chapter 3 was applied to Iraq war coalition casualty data. A 3σ control chart was used to detect a change in daily casualties, and then the change-point estimator was applied. It was shown that an increase in daily coalition casualties coincided with the start of the Muslim holy month of Ramadan. The significance of the example is while most of the literature deals with examples from manufacturing settings, this example shows these methods can be applied to a non-traditional, non-manufacturing process.

5.2 Future Research

Although a general methodology has been developed for estimating a step change in both the mean and overdispersion parameters of an overdispersed count process, there are areas of future research. For instance, in this thesis, the two-parameter gamma-Poisson mixture model was used to model overdispersed count data. It was assumed that both parameters were known for the in-control process, but that the process went out-of-control due to a step change in only one of the parameters. While, these assumptions allowed a foundation to be laid, other cases can be considered.

The case where neither the in-control nor out-of control parameter values are known would be a very valuable addition to this research. While any established process would have known or sufficiently estimated in-control parameter values, it is possible a new process could benefit from this kind of analysis. For example, at the start of the Iraq war, there may not have been any historical data. Certainly estimating a change in a process where none of the parameters are known, *a priori*, is important.

In addition, the assumption of a sustained step change can be too simplistic in some instances. The techniques in this thesis lend themselves well to a process where it is assumed that the type of change can be accurately modeled by a step function. However, if the type of change is not known before hand, then the procedure proposed by Perry, Pignatiello, and Simpson [14] may be a better alternative.

In this thesis, it was assumed that the variability of the Poisson rate parameter, λ, was modeled more accurately by allowing it to be modeled as the flexible, two-parameter gamma probability distribution. However, it may be the case that a set of covariates (predictors) do well in explaining some of the variability in λ. A change-point model that incorporates covariates may have the potential to model a wide range of process behaviors. For example, step changes, linear trends, and cyclical changes.

Last, this research proposes that the negative binomial distribution is an adequate model to use for all count data processes since the Poisson distribution is just a limiting form of the negative binomial distribution. However, the one-parameter Poisson distribution may be preferred in cases where it can be used. Hence, it may be worth characterizing for exactly what values of various combinations of the negative binomial parameters that the Poisson distribution is an adequate model to use in place of the negative binomial.

Appendix A: Simulation Data for Step Change in Mean

Table 22: Performance of MLE for $\alpha = 1,\ \nu_0 = 1,\ \tau = 50,$ and $N = 10,000$

ν_a	% Increase	ARL	RMS($\hat{\tau}_\nu$)	$\bar{\hat{\tau}}_\nu$	Std.error($\hat{\tau}_\nu$)
1.2	20	38.20	50.42	82.57	0.385
1.4	40	25.66	30.89	69.03	0.243
1.6	60	18.45	20.97	61.82	0.173
1.8	80	14.22	15.46	57.93	0.133
2	100	11.47	12.00	55.43	0.107
2.2	120	9.44	10.10	53.90	0.093
2.4	140	8.03	8.70	53.06	0.082
2.6	160	7.16	7.76	52.38	0.074
2.8	180	6.23	7.08	51.87	0.068
3	200	5.60	6.61	51.46	0.065
4	300	3.86	5.27	50.59	0.052
5	400	2.98	4.72	50.22	0.047
6	500	2.54	4.10	50.12	0.041
7	600	2.25	3.99	49.97	0.040
8	700	2.06	3.60	49.98	0.036
9	800	1.87	3.83	49.83	0.038
10	900	1.77	3.49	49.84	0.035
11	1000	1.70	3.23	49.86	0.032

Table 23: Performance of MLE for $\alpha = 1,\ \nu_0 = 5,\ \tau = 50,$ and $N = 10,000$

ν_a	% Increase	ARL	RMS($\hat{\tau}_\nu$)	$\bar{\hat{\tau}}_\nu$	Std.error($\hat{\tau}_\nu$)
6	20	28.90	37.49	73.50	0.292
7	40	18.65	21.99	62.62	0.180
8	60	13.41	14.78	57.41	0.128
9	80	10.12	11.23	54.64	0.102
10	100	8.16	9.07	53.35	0.084
11	120	6.80	7.96	52.21	0.076
12	140	5.74	7.08	51.56	0.069
13	160	5.06	6.40	51.30	0.063
14	180	4.58	5.96	51.04	0.059
15	200	4.05	5.93	50.70	0.059
20	300	2.95	4.52	50.23	0.045
25	400	2.37	4.33	50.01	0.043
30	500	2.05	4.06	49.88	0.041
35	600	1.84	4.15	49.78	0.042
40	700	1.73	3.28	49.86	0.033
45	800	1.63	3.55	49.81	0.035
50	900	1.56	3.57	49.82	0.036
55	1000	1.47	3.27	49.80	0.033

Table 24: Performance of MLE for $\alpha = 1,\ \nu_0 = 20,\ \tau = 50,$ and $N = 10,000$

ν_a	% Increase	ARL	RMS($\hat{\tau}_\nu$)	$\bar{\hat{\tau}}_\nu$	Std.error($\hat{\tau}_\nu$)
24	20	28.52	36.64	72.91	0.286
28	40	17.86	20.68	61.66	0.171
32	60	12.52	14.00	56.84	0.122
36	80	9.48	10.45	54.21	0.096
40	100	7.57	8.72	52.80	0.083
44	120	6.19	7.55	51.93	0.073
48	140	5.52	6.60	51.57	0.064
52	160	4.85	6.10	51.19	0.060
56	180	4.29	5.91	50.85	0.059
60	200	3.90	5.37	50.66	0.053
80	300	2.76	4.80	50.09	0.048
100	400	2.26	4.55	49.89	0.046
120	500	1.99	3.82	49.93	0.038
140	600	1.79	3.76	49.83	0.038
160	700	1.67	3.40	49.87	0.034
180	800	1.58	3.90	49.72	0.039
200	900	1.50	2.98	49.84	0.030
220	1000	1.45	2.90	49.84	0.029

Table 25: Performance of MLE for $\alpha = 5,\ \nu_0 = 1,\ \tau = 50,$ and $N = 10,000$

ν_a	% Increase	ARL	RMS($\hat{\tau}_\nu$)	$\bar{\hat{\tau}}_\nu$	Std.error($\hat{\tau}_\nu$)
1.2	20	59.63	69.91	95.41	0.532
1.4	40	35.53	31.62	68.00	0.260
1.6	60	23.10	17.99	58.16	0.160
1.8	80	16.20	11.95	54.42	0.111
2	100	12.29	9.13	52.63	0.088
2.2	120	9.33	7.42	51.61	0.072
2.4	140	7.60	6.37	50.97	0.063
2.6	160	6.40	5.68	50.58	0.057
2.8	180	5.40	5.25	50.27	0.053
3	200	4.66	5.09	50.04	0.051
4	300	2.73	4.10	49.65	0.041
5	400	1.99	3.79	49.57	0.038
6	500	1.63	3.47	49.58	0.035
7	600	1.43	3.23	49.61	0.032
8	700	1.31	2.76	49.66	0.027
9	800	1.23	2.70	49.68	0.027
10	900	1.17	2.32	49.74	0.023
11	1000	1.12	2.01	49.78	0.020

Table 26: Performance of MLE for $\alpha = 5,\ \nu_0 = 5,\ \tau = 50,$ and $N = 10,000$

ν_a	% Increase	ARL	RMS($\hat{\tau}_\nu$)	$\bar{\hat{\tau}}_\nu$	Std.error($\hat{\tau}_\nu$)
6	20	39.01	36.96	70.98	0.304
7	40	19.45	14.16	55.62	0.130
8	60	11.39	8.58	51.83	0.084
9	80	7.36	6.22	50.67	0.062
10	100	5.28	5.59	50.03	0.056
11	120	4.12	4.75	49.92	0.048
12	140	3.31	4.40	49.71	0.044
13	160	2.78	4.40	49.57	0.044
14	180	2.41	4.00	49.61	0.040
15	200	2.15	4.05	49.50	0.040
20	300	1.49	3.62	49.55	0.036
25	400	1.25	2.57	49.70	0.026
30	500	1.14	2.46	49.71	0.024
35	600	1.08	1.93	49.81	0.019
40	700	1.06	1.62	49.84	0.016
45	800	1.03	1.39	49.89	0.014
50	900	1.02	0.96	49.93	0.010
55	1000	1.02	1.13	49.92	0.011

Table 27: Performance of MLE for $\alpha = 5,\ \nu_0 = 20,\ \tau = 50,$ and $N = 10,000$

ν_a	% Increase	ARL	RMS($\hat{\tau}_\nu$)	$\bar{\hat{\tau}}_\nu$	Std.error($\hat{\tau}_\nu$)
24	20	35.10	27.55	64.27	0.236
28	40	15.27	10.44	53.07	0.100
32	60	8.49	6.77	50.78	0.067
36	80	5.35	4.97	50.11	0.050
40	100	3.89	4.73	49.77	0.047
44	120	3.11	4.29	49.67	0.043
48	140	2.53	4.27	49.53	0.042
52	160	2.15	3.92	49.53	0.039
56	180	1.90	3.66	49.52	0.036
60	200	1.72	3.54	49.53	0.035
80	300	1.29	3.03	49.61	0.030
100	400	1.14	2.21	49.75	0.022
120	500	1.08	1.93	49.82	0.019
140	600	1.04	1.41	49.88	0.014
160	700	1.03	1.30	49.91	0.013
180	800	1.02	0.85	49.94	0.009
200	900	1.01	0.60	49.96	0.006
220	1000	1.01	0.39	49.98	0.004

Table 28: Performance of MLE for $\alpha = 10,\ \nu_0 = 1,\ \tau = 50,$ and $N = 10,000$

ν_a	% Increase	ARL	RMS($\hat{\tau}_\nu$)	$\bar{\hat{\tau}}_\nu$	Std.error($\hat{\tau}_\nu$)
1.2	20	80.75	86.03	105.81	0.655
1.4	40	47.12	34.74	68.55	0.294
1.6	60	29.33	18.31	58.14	0.164
1.8	80	20.04	12.07	53.93	0.114
2	100	14.68	8.91	52.31	0.086
2.2	120	10.94	7.26	51.35	0.071
2.4	140	8.70	6.38	50.65	0.063
2.6	160	7.11	5.35	50.47	0.053
2.8	180	5.76	4.99	50.18	0.050
3	200	4.88	4.83	49.96	0.048
4	300	2.70	3.84	49.65	0.038
5	400	1.91	3.38	49.61	0.034
6	500	1.54	3.14	49.62	0.031
7	600	1.33	3.02	49.60	0.030
8	700	1.22	2.42	49.71	0.024
9	800	1.15	2.16	49.77	0.022
10	900	1.10	1.98	49.79	0.020
11	1000	1.07	1.61	49.84	0.016

Table 29: Performance of MLE for $\alpha = 10,\ \nu_0 = 5,\ \tau = 50,$ and $N = 10,000$

ν_a	% Increase	ARL	RMS($\hat{\tau}_\nu$)	$\bar{\hat{\tau}}_\nu$	Std.error($\hat{\tau}_\nu$)
6	20	51.07	35.04	68.49	0.298
7	40	22.41	11.77	53.68	0.112
8	60	11.93	7.02	50.98	0.070
9	80	7.28	5.22	50.26	0.052
10	100	4.91	4.30	49.89	0.043
11	120	3.66	4.02	49.69	0.040
12	140	2.86	3.80	49.64	0.038
13	160	2.36	3.78	49.53	0.038
14	180	2.07	3.35	49.60	0.033
15	200	1.79	3.37	49.54	0.033
20	300	1.26	2.71	49.65	0.027
25	400	1.11	2.08	49.78	0.021
30	500	1.04	1.30	49.89	0.013
35	600	1.02	1.09	49.92	0.011
40	700	1.01	0.79	49.95	0.008
45	800	1.00	0.51	49.97	0.005
50	900	1.00	0.47	49.98	0.005
55	1000	1.00	0.70	49.98	0.007

Table 30: Performance of MLE for $\alpha = 10,\ \nu_0 = 20,\ \tau = 50,$ and $N = 10,000$

ν_a	% Increase	ARL	RMS($\hat{\tau}_\nu$)	$\bar{\hat{\tau}}_\nu$	Std.error($\hat{\tau}_\nu$)
24	20	34.94	20.11	58.28	0.183
28	40	12.80	7.68	51.05	0.076
32	60	6.41	5.33	49.94	0.053
36	80	3.97	4.24	49.67	0.042
40	100	2.75	3.72	49.58	0.037
44	120	2.13	3.39	49.57	0.034
48	140	1.77	3.11	49.59	0.031
52	160	1.54	3.14	49.57	0.031
56	180	1.38	2.99	49.60	0.030
60	200	1.28	2.74	49.64	0.027
80	300	1.07	1.75	49.80	0.017
100	400	1.02	1.32	49.90	0.013
120	500	1.01	0.75	49.95	0.008
140	600	1.00	0.35	49.98	0.004
160	700	1.00	0.29	49.99	0.003
180	800	1.00	0.18	49.99	0.002
200	900	1.00	0.06	50.00	0.001
220	1000	1.00	0.04	50.00	0.000

Table 31: Performance of MLE for $\alpha = 50,\ \nu_0 = 1,\ \tau = 50,$ and $N = 10,000$

ν_a	% Increase	ARL	RMS($\hat{\tau}_\nu$)	$\bar{\hat{\tau}}_\nu$	Std.error($\hat{\tau}_\nu$)
1.2	20	117.43	103.96	115.84	0.805
1.4	40	64.79	35.47	67.90	0.306
1.6	60	39.11	18.26	57.18	0.168
1.8	80	25.67	11.56	53.28	0.111
2	100	17.80	8.47	51.84	0.083
2.2	120	13.02	6.53	51.05	0.064
2.4	140	9.96	5.52	50.57	0.055
2.6	160	7.81	5.04	50.29	0.050
2.8	180	6.35	4.68	50.06	0.047
3	200	5.26	4.71	49.87	0.047
4	300	2.68	3.53	49.65	0.035
5	400	1.81	3.14	49.60	0.031
6	500	1.44	2.93	49.65	0.029
7	600	1.23	2.61	49.67	0.026
8	700	1.13	1.87	49.78	0.019
9	800	1.07	1.60	49.83	0.016
10	900	1.04	1.53	49.86	0.015
11	1000	1.02	1.29	49.89	0.013

Table 32: Performance of MLE for $\alpha = 50,\ \nu_0 = 5,\ \tau = 50,$ and $N = 10,000$

ν_a	% Increase	ARL	RMS($\hat{\tau}_\nu$)	$\bar{\hat{\tau}}_\nu$	Std.error($\hat{\tau}_\nu$)
6	20	75.77	30.75	63.39	0.277
7	40	27.62	9.54	51.92	0.094
8	60	12.84	5.63	50.42	0.056
9	80	7.12	4.27	49.92	0.043
10	100	4.45	3.54	49.77	0.035
11	120	3.14	3.36	49.65	0.033
12	140	2.38	3.26	49.60	0.032
13	160	1.91	3.11	49.59	0.031
14	180	1.63	3.01	49.59	0.030
15	200	1.41	2.67	49.64	0.027
20	300	1.08	1.61	49.82	0.016
25	400	1.01	0.95	49.93	0.010
30	500	1.00	0.33	49.98	0.003
35	600	1.00	0.08	50.00	0.001
40	700	1.00	0.05	50.00	0.001
45	800	1.00	0.02	50.00	0.000
50	900	1.00	0.00	50.00	0.000
55	1000	1.00	0.01	50.00	0.000

Table 33: Performance of MLE for $\alpha = 50,\ \nu_0 = 20,\ \tau = 50,$ and $N = 10,000$

ν_a	% Increase	ARL	RMS($\hat{\tau}_\nu$)	$\bar{\hat{\tau}}_\nu$	Std.error($\hat{\tau}_\nu$)
24	20	28.56	10.87	52.56	0.106
28	40	7.76	4.54	49.94	0.045
32	60	3.33	3.50	49.61	0.035
36	80	1.96	3.13	49.56	0.031
40	100	1.45	2.97	49.60	0.029
44	120	1.21	2.57	49.66	0.026
48	140	1.10	2.37	49.73	0.024
52	160	1.04	1.63	49.84	0.016
56	180	1.02	1.10	49.91	0.011
60	200	1.01	0.76	49.94	0.008
80	300	1.00	0.12	49.99	0.001
100	400	1.00	0.05	50.00	0.001
120	500	1.00	0.00	50.00	0.000
140	600	1.00	0.00	50.00	0.000
160	700	1.00	0.00	50.00	0.000
180	800	1.00	0.00	50.00	0.000
200	900	1.00	0.00	50.00	0.000
220	1000	1.00	0.00	50.00	0.000

Appendix B: Cardinality and Coverage Surface Plots for Step Change in Mean

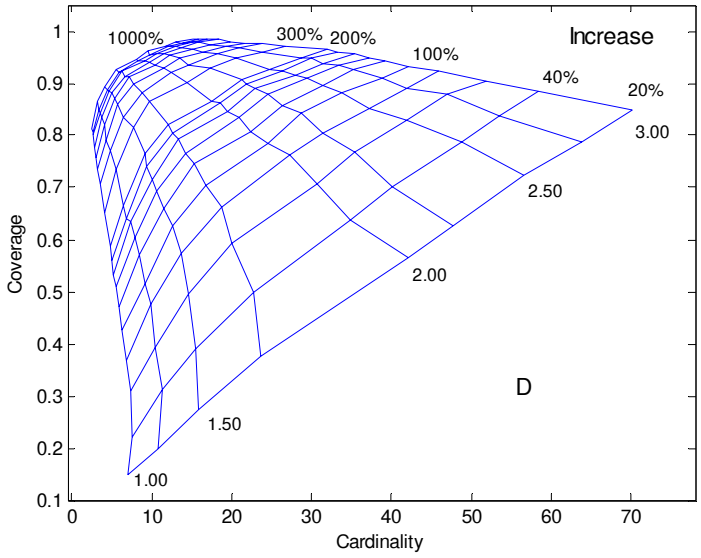

Figure 17: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Set Cardinality, Coverage, Percent Increase from ν_0, and Reference Value D. $\alpha = 1,\ \nu_0 = 1,\ \tau = 50,$ and $N = 10,000.$

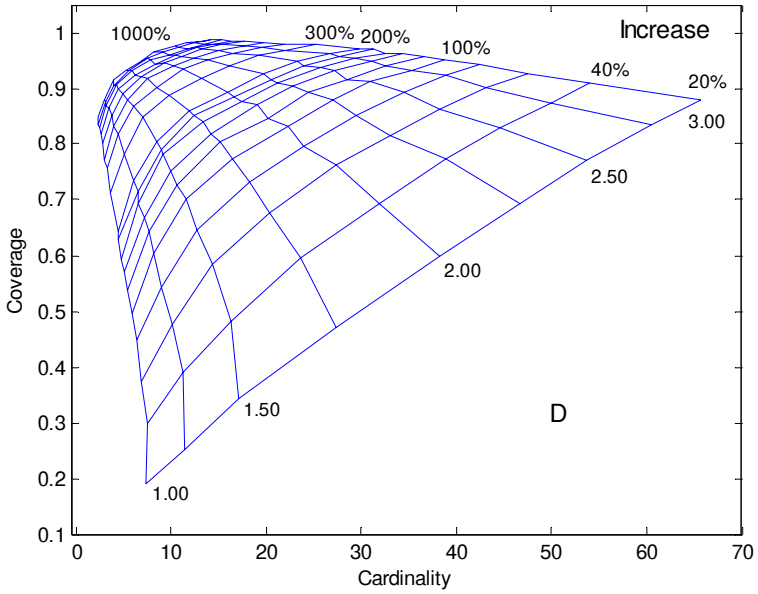

Figure 18: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Set Cardinality, Coverage, Percent Increase from ν_0, and Reference Value D. $\alpha = 1,\ \nu_0 = 5,\ \tau = 50,$ and $N = 10,000.$

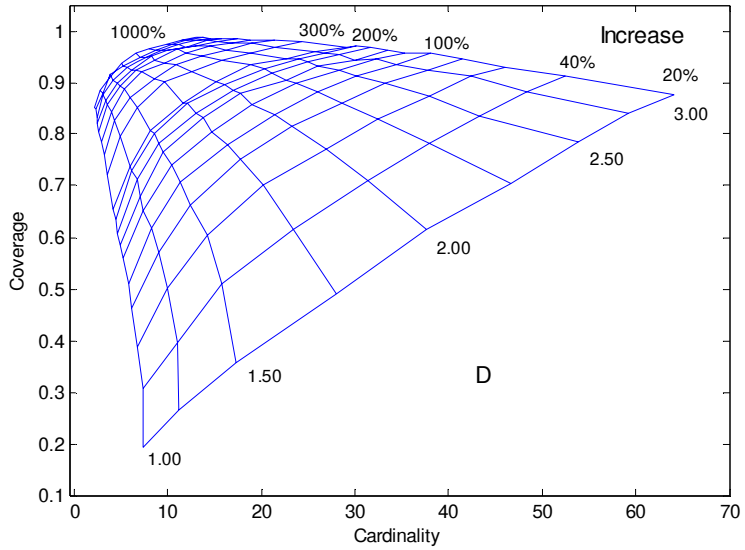

Figure 19: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Set Cardinality, Coverage, Percent Increase from ν_0, and Reference Value D. $\alpha = 1,\ \nu_0 = 20,\ \tau = 50,$ and $N = 10,000.$

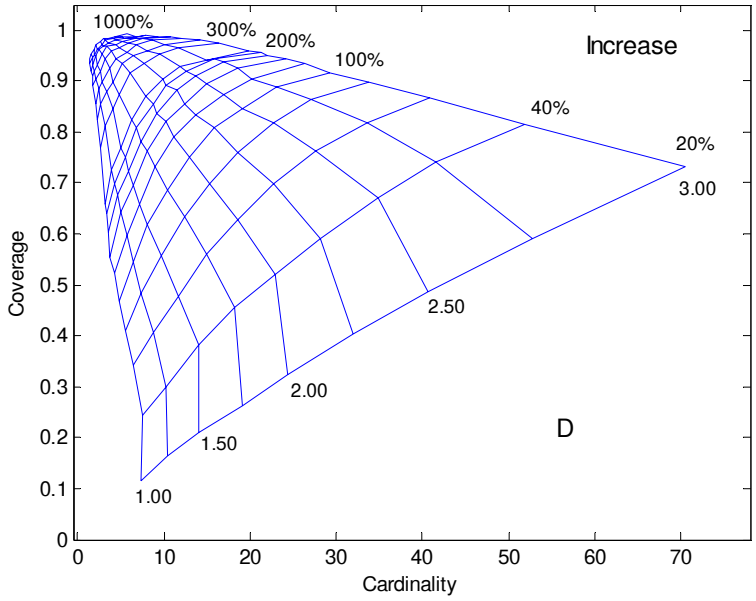

Figure 20: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Set Cardinality, Coverage, Percent Increase from ν_0, and Reference Value D. $\alpha = 5,\ \nu_0 = 1,\ \tau = 50,$ and $N = 10,000.$

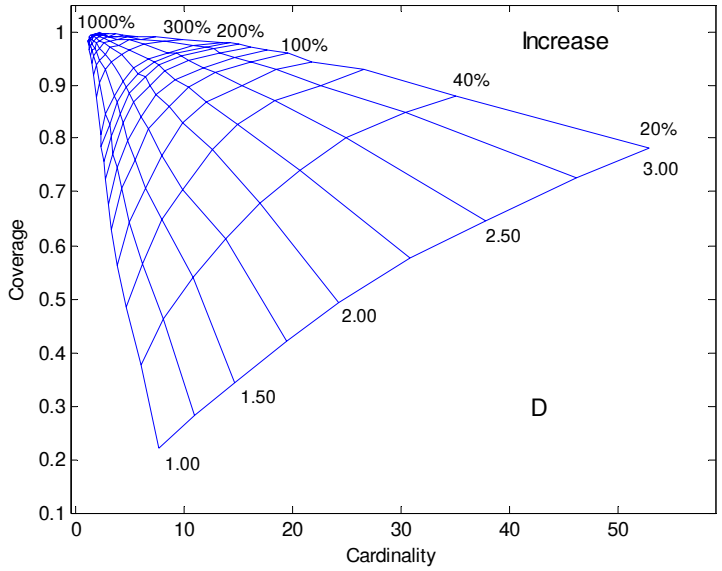

Figure 21: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Set Cardinality, Coverage, Percent Increase from ν_0, and Reference Value D. $\alpha = 5,\ \nu_0$=5, τ=50, and $N = 10,000.$

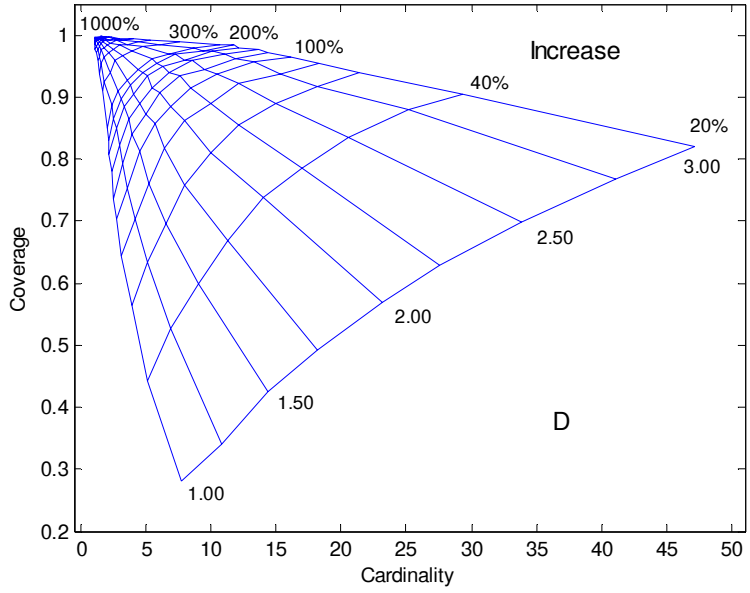

Figure 22: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Set Cardinality, Coverage, Percent Increase from ν_0, and Reference Value D. $\alpha = 5,\ \nu_0$=20, τ=50, and $N = 10,000.$

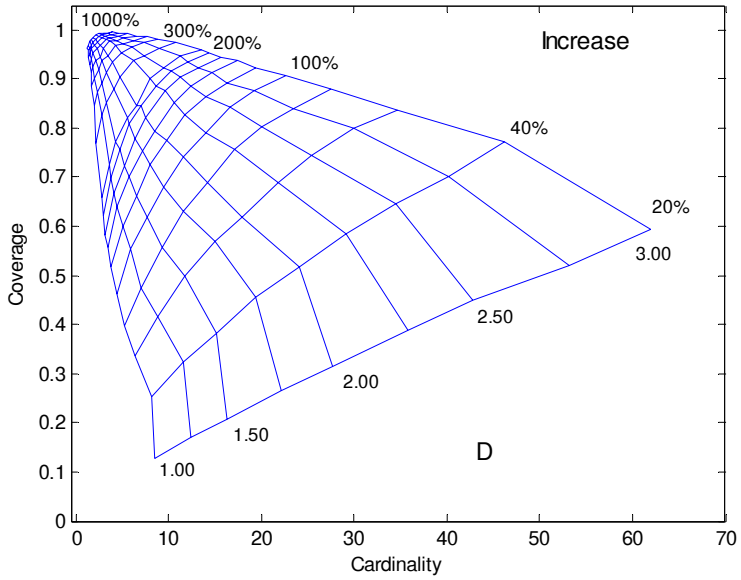

Figure 23: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Set Cardinality, Coverage, Percent Increase from ν_0, and Reference Value D. $\alpha = 10,\ \nu_0 = 1,\ \tau = 50,$ and $N = 10,000.$

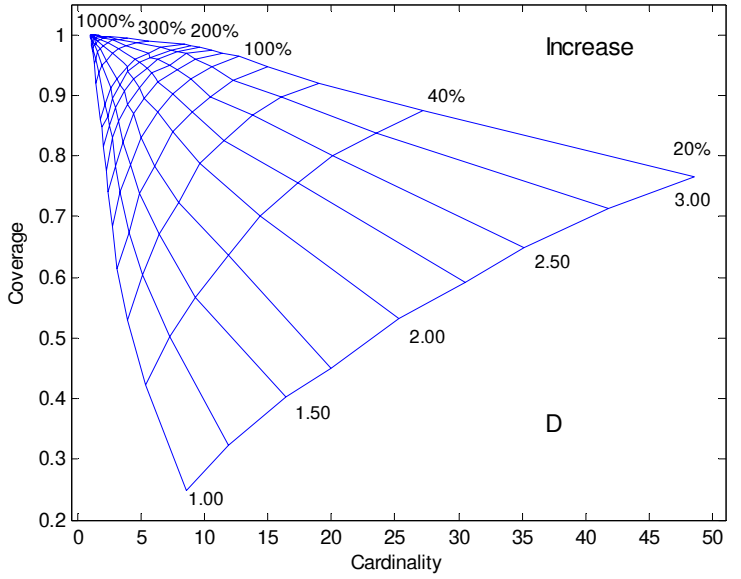

Figure 24: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Set Cardinality, Coverage, Percent Increase from ν_0, and Reference Value D. $\alpha = 10,\ \nu_0 = 5,\ \tau = 50,$ and $N = 10,000.$

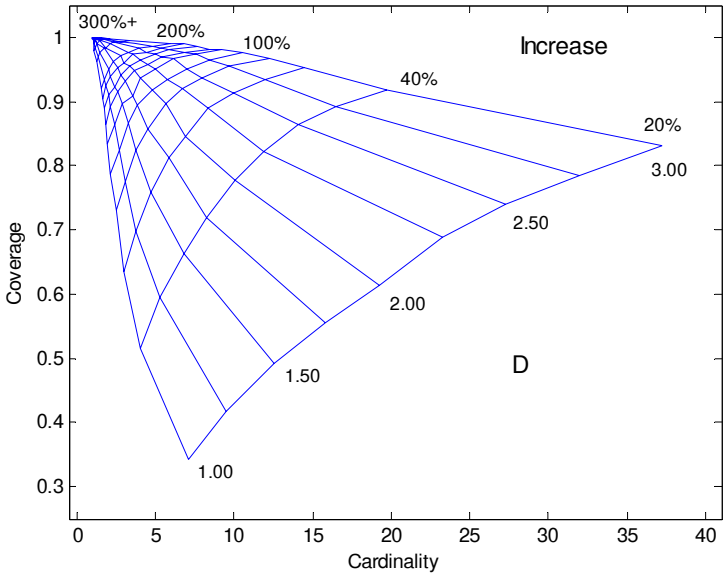

Figure 25: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Set Cardinality, Coverage, Percent Increase from ν_0, and Reference Value D. $\alpha = 10,\ \nu_0 = 20,\ \tau = 50,$ and $N = 10,000.$

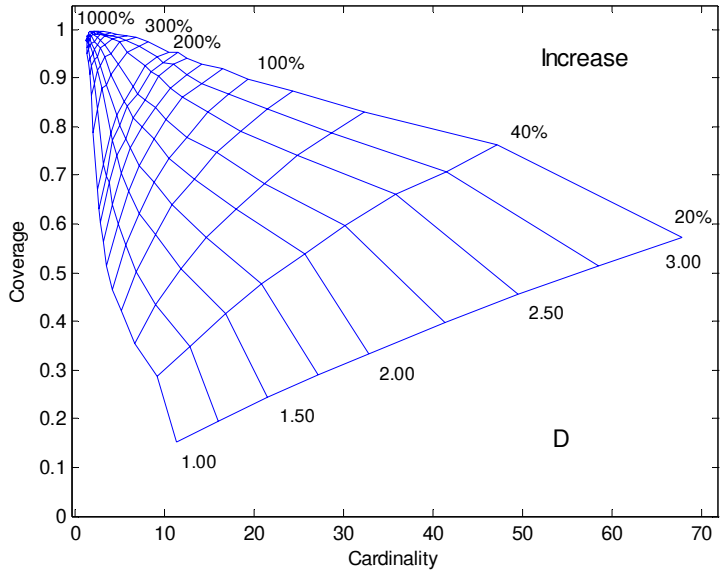

Figure 26: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Set Cardinality, Coverage, Percent Increase from ν_0, and Reference Value D. $\alpha = 50,\ \nu_0 = 1,\ \tau = 50,$ and $N = 10,000.$

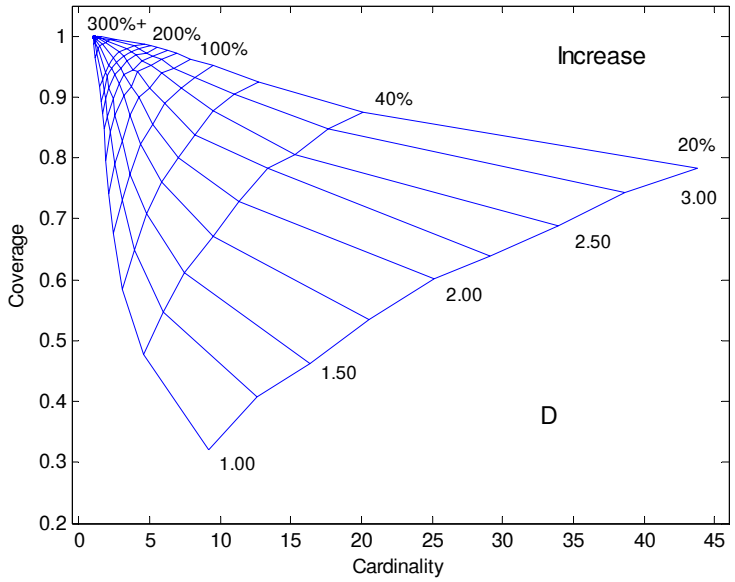

Figure 27: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Set Cardinality, Coverage, Percent Increase from ν_0, and Reference Value D. $\alpha = 50$, ν_0=5, τ=50, and $N = 10,000$.

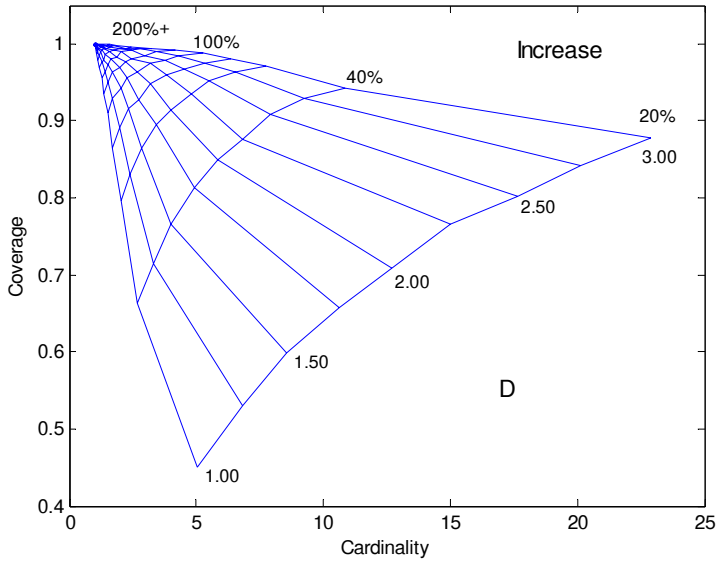

Figure 28: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Set Cardinality, Coverage, Percent Increase from ν_0, and Reference Value D. $\alpha = 50$, ν_0=20, τ=50, and $N = 10,000$.

Appendix C: Simulation Data for Step Change in Overdispersion Parameter

Table 34: Performance of MLE for $\nu = 5,\ \alpha_0 = 1,\ \tau$=50, and $N = 10,000$

α_a	% Decrease	ARL	RMS($\hat{\tau}_\alpha$)	$\bar{\hat{\tau}}_\alpha$	Std.error($\bar{\hat{\tau}}_\alpha$)
0.9	10	181.60	235.80	210.09	1.731
0.8	20	97.94	117.52	127.45	0.884
0.7	30	59.05	62.93	89.00	0.494
0.6	40	38.45	36.12	69.73	0.303
0.5	50	26.64	22.19	59.73	0.200
0.4	60	18.84	14.33	53.83	0.138
0.3	70	13.77	9.95	50.64	0.099
0.2	80	10.04	7.05	49.54	0.070
0.1	90	7.38	5.97	48.63	0.058

Table 35: Performance of MLE for $\nu = 5,\ \alpha_0 = 5,\ \tau$=50, and $N = 10,000$

α_a	% Decrease	ARL	RMS($\hat{\tau}_\alpha$)	$\bar{\hat{\tau}}_\alpha$	Std.error($\bar{\hat{\tau}}_\alpha$)
4.5	10	218.20	290.90	248.63	2.125
4	20	140.78	179.69	171.56	1.323
3.5	30	89.11	104.73	120.50	0.774
3	40	60.38	65.26	91.97	0.500
2.5	50	40.36	39.97	72.80	0.328
2	60	27.86	25.10	62.02	0.220
1.5	70	18.77	15.03	55.04	0.142
1	80	12.22	9.54	51.54	0.094
0.5	90	7.36	6.21	49.78	0.062

Table 36: Performance of MLE for $\nu = 5,\ \alpha_0 = 10,\ \tau$=50, and $N = 10,000$

α_a	% Decrease	ARL	RMS($\hat{\tau}_\alpha$)	$\bar{\hat{\tau}}_\alpha$	Std.error($\bar{\hat{\tau}}_\alpha$)
9	10	246.97	327.69	276.17	2.371
8	20	175.21	226.82	205.20	1.654
7	30	124.48	153.54	155.15	1.119
6	40	85.02	98.24	115.80	0.730
5	50	59.27	63.04	90.23	0.485
4	60	40.55	39.48	72.81	0.322
3	70	26.70	23.51	60.72	0.209
2	80	16.92	13.65	53.77	0.131
1	90	9.26	7.75	50.18	0.078

Table 37: Performance of MLE for $\nu = 20,\ \alpha_0 = 1,\ \tau$=50, and $N = 10,000$

α_a	% Decrease	ARL	RMS($\hat{\tau}_\alpha$)	$\bar{\hat{\tau}}_\alpha$	Std.error($\bar{\hat{\tau}}_\alpha$)
0.9	10	151.83	198.58	184.50	1.461
0.8	20	76.51	89.16	109.05	0.668
0.7	30	43.77	46.07	77.89	0.367
0.6	40	27.73	25.38	63.51	0.215
0.5	50	18.87	15.70	56.31	0.144
0.4	60	13.30	10.27	52.61	0.099
0.3	70	9.46	7.22	50.53	0.072
0.2	80	6.77	5.68	49.51	0.057
0.1	90	4.76	4.57	48.98	0.045

Table 38: Performance of MLE for $\nu = 20,\ \alpha_0 = 5,\ \tau$=50, and $N = 10,000$

α_a	% Decrease	ARL	RMS($\hat{\tau}_\alpha$)	$\bar{\hat{\tau}}_\alpha$	Std.error($\bar{\hat{\tau}}_\alpha$)
4.5	10	177.80	235.03	209.76	1.724
4	20	93.53	114.08	126.01	0.851
3.5	30	55.23	61.16	88.53	0.475
3	40	35.67	36.15	70.57	0.297
2.5	50	23.46	21.45	60.60	0.186
2	60	15.53	13.51	54.73	0.127
1.5	70	10.52	9.17	52.00	0.090
1	80	6.77	6.23	50.51	0.062
0.5	90	3.99	4.11	49.86	0.041

Table 39: Performance of MLE for $\nu = 20,\ \alpha_0 = 10,\ \tau$=50, and $N = 10,000$

α_a	% Decrease	ARL	RMS($\hat{\tau}_\alpha$)	$\bar{\hat{\tau}}_\alpha$	Std.error($\bar{\hat{\tau}}_\alpha$)
9	10	192.08	252.41	223.78	1.831
8	20	112.28	140.27	144.19	1.039
7	30	66.85	76.22	99.57	0.579
6	40	43.35	45.26	77.37	0.360
5	50	28.19	26.67	64.21	0.226
4	60	19.11	16.79	57.04	0.152
3	70	12.62	10.87	52.88	0.105
2	80	8.02	7.47	50.93	0.074
1	90	4.44	4.76	50.01	0.048

Appendix D: Cardinality and Coverage Surface Plots for Step Change in Overdispersion Parameter

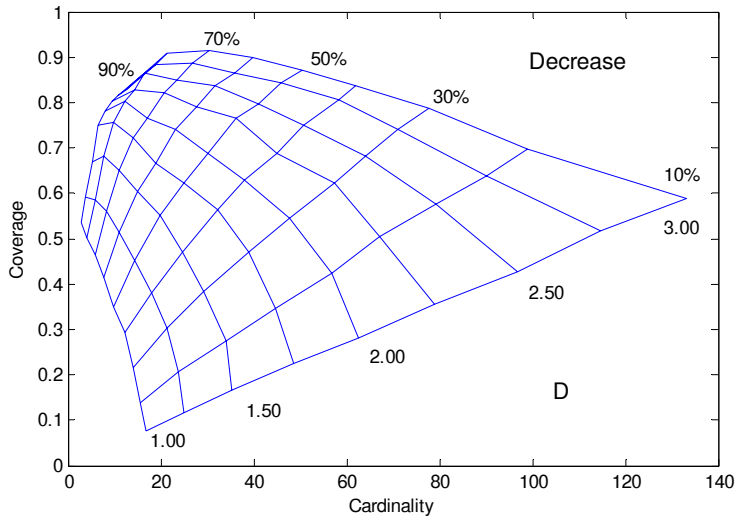

Figure 29: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Cardinality, Coverage, Percent Decrease from α_0, and Reference Value D. ν=5, α_0=1, τ=50, and $N = 10,000$.

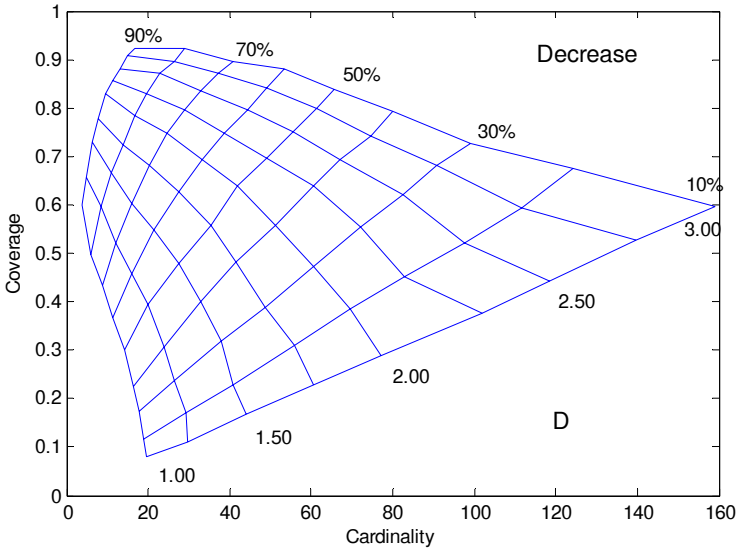

Figure 30: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Cardinality, Coverage, Percent Decrease from α_0, and Reference Value D. ν=5, α_0=5, τ=50, and $N = 10,000$.

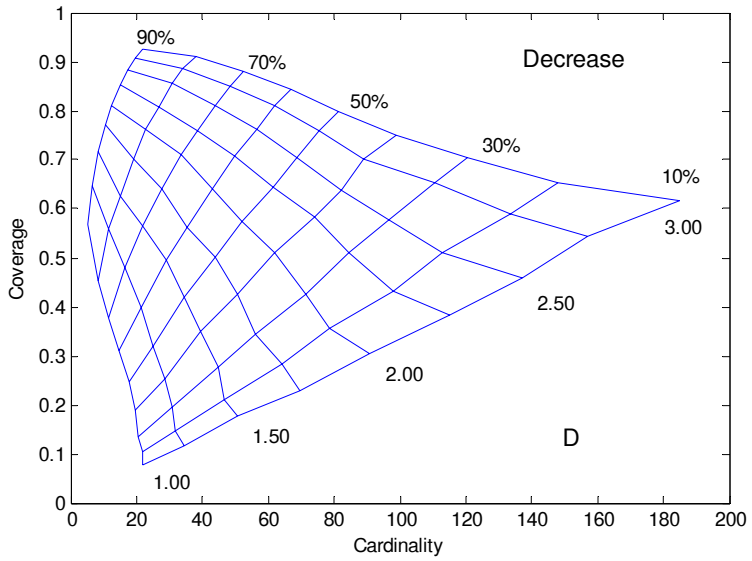

Figure 31: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Cardinality, Coverage, Percent Decrease from α_0, and Reference Value D. ν=5, α_0=10, τ=50, and $N = 10,000$.

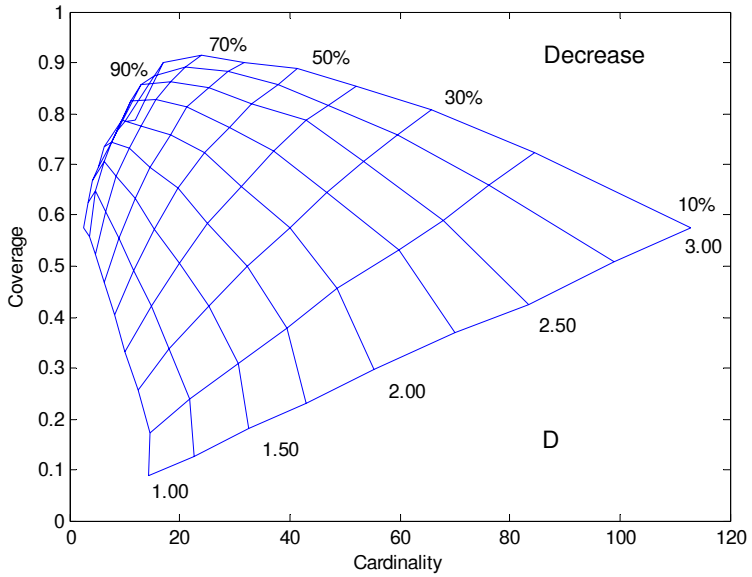

Figure 32: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Cardinality, Coverage, Percent Decrease from α_0, and Reference Value D. ν=20, α_0=1, τ=50, and $N = 10,000$.

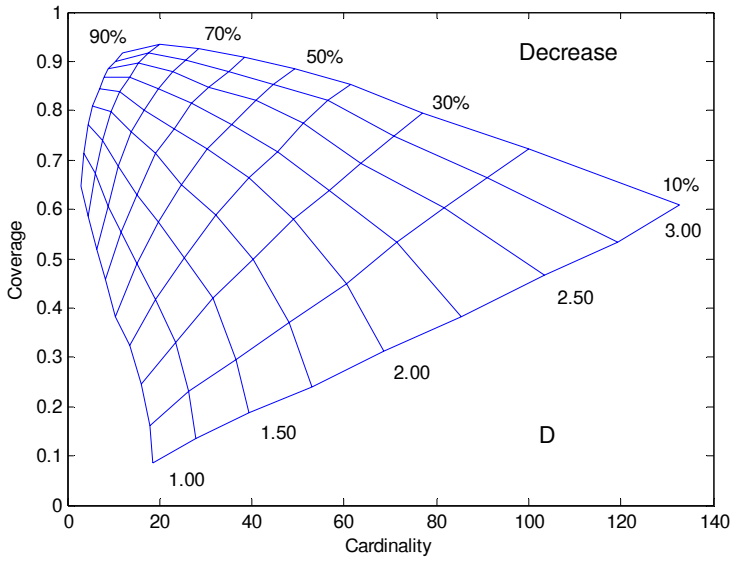

Figure 33: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Cardinality, Coverage, Percent Decrease from α_0, and Reference Value D. ν=20, α_0=5, τ=50, and $N = 10,000$.

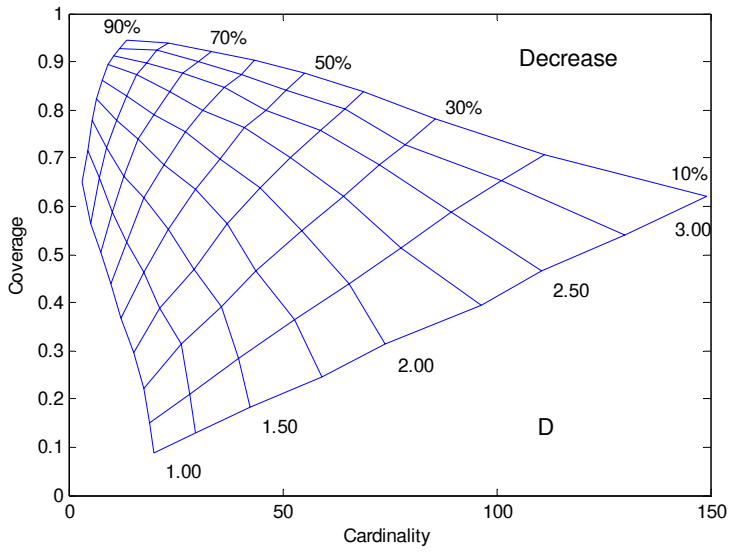

Figure 34: Surface Plot Obtained from Confidence Set Estimator Showing Estimated Relationships Between Cardinality, Coverage, Percent Decrease from α_0, and Reference Value D. ν=20, α_0=10, τ=50, and $N = 10,000$.

Bibliography

[1] Basseville, M., & Nikiforov, I. V. (1993). *Detection and Theory of Abrupt Changes: Theory and Application*. New Jersey: Prentice Hall.

[2] Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations. *Journal of Royal Statistical Society B, 26*(2), 211-243.

[3] Burden, R. L., & Faires, J. D. (2005). *Numerical Analysis* (8th ed.). California: Brooks/Cole.

[4] Cameron, A. C., & Trivedi, P. K. (1998). *Regression Analysis of Count Data*. New York: Cambridge University Press.

[5] Haight, F. A. (1967). *Handbook of the Poisson Distribution*. New York: John Wiley & Sons.

[6] Hawkins, D. M., & Olwell, D. H. (1998). *Cumulative Sum Charts and Charting for Quality Improvement*. New York: Springer-Verlag.

[7] Hinkley, D. V. (1970). Inference About the Change-Point in a Sequence of Random Variables. *Biometrika, 57*(1), 1-17.

[8] Khoo, M. B. C. (2005). Determining the Time of a Permanent Shift in the Process Mean of CUSUM Control Charts. *Quality Engineering, 17*(1), 87-93.

[9] Montgomery, D. C. (2005). *Introduction to Statistical Quality Control* (5th ed.). New York: John Wiley & Sons.

[10] Nishina, K. (1992). Comparison of Control Charts from the Viewpoint of Change-Point Estimation. *Quality and Reliability Engineering International, 8*(6), 537-541.

[11] Park, J., & Park, S. (2004). Estimation of the Change Point in the X over-bar and S Control Charts. *Communications in Statistics Part B: Simulation and Computation, 33*(4), 1115-1132.

[12] Perry, M. B., & Pignatiello, J. J. (2006). Estimating the Change Point of a Normal Process Mean with a Linear Trend Disturbance in SPC. *Quality Technology & Quantitative Management, 3*(3), 325-334.

[13] Perry, M. B., & Pignatiello, J. J. (2005). Estimation of the Change Point of the Process Fraction Nonconforming in SPC Applications. *International Journal of Reliability, Quality and Safety Engineering, 12*(2), 95-110.

[14] Perry, M. B., Pignatiello, J. J., & Simpson, J. R. (2007). Change Point Estimation for Monotonically Changing Poisson Rates in SPC. *International Journal of Production Research, 45*(8), 1791-1813.

[15] Perry, M. B., Pignatiello, J. J., & Simpson, J. R. (2006). Estimating the Change Point of a Poisson Rate Parameter with a Linear Trend Disturbance. *Quality and Reliability Engineering International, 22*(4), 371-384.

[16] Pignatiello, J. J., & Samuel, T. R. (2001). Estimation of the Change Point of a Normal Process Mean in SPC Applications. *Journal of Quality Technology, 33*(1), 82-95.

[17] Pignatiello, J. J., & Samuel, T. R. (2001). Identifying the Time of a Step-Change in the Process Fraction Nonconforming. *Quality Engineering, 13*(4), 357-365.

[18] Roberts, S. W. (1959). Control Chart Tests Based on Geometric Moving Averages. *Technometrics*,1, 239-250.

[19] Ryan, T. P. (2000). *Statistical Methods for Quality Improvement* (2nd ed.). New York: John Wiley & Sons.

[20] Samuel, T. R., & Pignatiello, J. J. (1998). Identifying the Time of a Change in a Poisson Rate Parameter. *Quality Engineering, 10*(4), 673-681.

[21] Samuel, T. R., Pignatiello, J. J., & Calvin, J. A. (1998). Identifying the Time of a Step Change with X Control Charts. *Quality Engineering, 10*(3), 521-527.

[22] Wackerly, D. D., Mendenhall, W., III., & Scheaffer, R. L. (2002). *Mathematical Statistics with Applications* (6th ed.). Pacific Grove CA: Duxbury.

Vita

Captain Brian A. Wilken graduated from Kankakee High School, Kankakee, IL. Shortly after graduation he entered Kankakee Community College and later transferred to Olivet Nazarene University, Bourbonnais, IL. In January 2002 he graduated from Olivet Nazarene University with a Bachelor of Arts in secondary mathematics education. In August 2002 he was commissioned as an Air Force officer through Officer Training School, Maxwell AFB, Alabama.

His first assignment was Detachment 1, 31st Test and Evaluation Squadron, Kirtland Air Force Base, NM as a weapons systems analyst. While there his primary role was as an analyst on the F/A-22 test team during Initial Operational Test & Evaluation (IOT&E). In August 2005 he entered the Air Force Institute of Technology to study Operations Research with an emphasis in applied statistics. Upon his March 2007 graduation, he will work at the Management Sciences Division, Headquarters Air Force Material Command (HQ AFMC/A8S), Wright-Patterson AFB.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From – To)*
23-03-2007	Master's Thesis	Sep 2005 - Mar 2007

4. TITLE AND SUBTITLE
Change-Point Methods for Overdispersed Count Data

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Wilken, Brian A., Captain, USAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)
Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Street, Building 642
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT/GOR/ENS/07-26

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
N/A

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

A control chart is often used to detect a change in a process. Following a control chart signal, knowledge of the time and magnitude of the change would simplify the search for and identification of the assignable cause. In this research, emphasis is placed on count processes where overdispersion has occurred. Overdispersion is common in practice and occurs when the observed variance is larger than the theoretical variance of the assumed model. Although the Poisson model is often used to model count data, the two-parameter gamma-Poisson mixture parameterization of the negative binomial distribution is often a more adequate model for overdispersed count data. In this research effort, maximum likelihood estimators for the time of a step change in each of the parameters of the gamma-Poisson mixture model are derived. Monte Carlo simulation is used to evaluate the root mean square error performance of these estimators to determine their utility in estimating the change point, following a control chart signal. Results show that the estimators provide process engineers with accurate and useful estimates for the time of step change. In addition, an approach for estimating a confidence set for the process change point will be presented.

15. SUBJECT TERMS
Statistical Process Control; Quality Control; Count Data; CUSUM Control Chart; Shewhart Control Chart; Maximum Likelihood Estimation; Change Point Estimation; Confidence Set Estimation; Negative Binomial;

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Marcus B. Perry, PhD
U	U	U	UU	110	19b. TELEPHONE NUMBER *(Include area code)* (937) 255-3636 x4588 e-mail: Marcus.Perry@AFIT.edu

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39-18